POPULATION AND GENOMIC VARIATION OF IMMUNE GENES IN CHICKEN

Tim Downing

A thesis submitted to the University of Dublin

for the degree of Doctor of Philosophy

Smurfit Institute of Genetics

Trinity College

University of Dublin

January 2010

Declaration

I hereby certify that this thesis has not been previously submitted for examination to this or any other university. The work described herein has been carried out by the author alone, except where otherwise stated.

This thesis may be made available for consultation within the university library and may be photocopied or loaned to other libraries for the purposes of consultation.

[redacted]

Summary

Achieving an understanding of the processes shaping diversity at chicken immune genes illuminates their population history, relevance to disease and mechanisms of evolution. Functional variation at genes that determine the resistance and susceptibility of chickens to infectious diseases can be identified by a combination of genomic surveys of variability and sequencing in diverse populations of modern birds. In this context, the work presented here describes the design and implementation of population and evolutionary genetic tests used to analyse the myriad effects of demographic history, pathogen-driven selection and functional constraint in the chicken immunome.

Two genomic approaches evaluated diversity in chicken immune genes. The first of these performed tests for adaptive evolution on a set of chicken and zebra finch orthologous genes whose functions were assigned from their human orthologs. As implied by other genome-wide studies, there was evidence that immune genes were under positive selection since the divergence of the common ancestor of chicken and zebra finch. The second genome-based strategy identified polymorphic sites in expressed sequence tags in previously sequenced chicken libraries. A database for these was created and corroboration of candidate variable sites was conducted on two immune genes subsequently resequenced.

Global chicken samples from diverse populations were collected, as were red, grey, Ceylon and green jungle fowl in addition to grey francolin and bamboo partridge. In order to investigate the chicken's complex population history of multiple origins and extensive migration, two chicken genes encoding interleukin-1β and interferon-γ from a literature database of genes associated with resistance or susceptibility to disease were sequenced in these populations. Variation at these genes exhibited contrasting features, but was nonetheless elevated.

Interspecies tests for positive selection were carried out on the dataset of chicken and zebra finch gene pairs to identify functionally important genes. One of these genes, the interleukin-4 receptor alpha chain, was sequenced in the above samples and in a set of commercial chickens as well. The pattern of diversity at this gene was balanced around two key amino acid-altering mutations that were present in all populations, including broilers. These samples were investigated further at the lysozyme gene, a key innate immune gene. This gene also displayed high variability centred on two amino acids close to the catalytic sites of the enzyme. This pattern of elevated functional diversity in all chicken populations indicated that extensive admixture and migration of chicken populations occurred after an initial series of domestication events.

Using newly developed sequencing technology, a set of cytokine and toll-like receptor loci were amplified in a smaller set of broilers, heritage chickens and red, grey, Ceylon and green jungle fowl. These gene classes had differing selection signatures that may be related to their separate functional roles: toll-like receptors interact directly with a limited number of pathogen molecules and so must adapt swiftly and directionally to their evolution. Cytokines are central signalling molecules that indirectly respond to many infectious challenges and as a result, they appear to be subject to frequency-dependent selection. There is evidence for this pattern in other vertebrate species. Analysis of bird population and species differentiation suggested no evidence of an ancient separation of chicken and red jungle fowl genetically. This demonstrated that chicken and wild jungle fowl have historically bred together. At two instances, variation was shared between chicken, red and grey jungle fowl, indicating that the domestic chicken may have more than one genetic contribution from grey jungle fowl. The sequencing of the toll-like receptors and cytokines continued the trend of high diversity and low population differentiation, confirming the chicken's population history has many origins, and may have been enhanced by human trade, assertions supported by other investigations of chicken genes.

The population and genomic approaches implemented here that determined the level of variation within the domestic chicken and its relationship with wild jungle fowl, as well as identifying pathogen-associated mutations at immune genes in chicken and jungle fowl will be useful for breeding birds less susceptible to infections, in the development of novel therapeutics for resisting diseases in birds, and in preventing transmission of infections to human from avian sources.

Acknowledgements

I would like to thank Dan for his guidance, insights and ability to isolate the key features in our ornithimmunomics. Working with Dan was a pleasure – I think this thesis is a testament to that! I would like to thank David, Andrew and Cliona for their contributions to this thesis. David for getting the work up and running, and for the bioinformatic ideas and strategies. Andrew for your patience in remodelling my muddled ideas to ones that might work, reading my countless scribbles, Dr Pedant and the conservations on what does it really mean. Cliona for her limitless enthusiasm and drive that led to us doing more than just sequencing to see what was there.

Thanks to everyone in the lab for their help and time: Lilian, Emma, Valeria, Yonas, Frauke, Russell, Matthew, Brian, Ceiridwen, Caitriona, Helge, Fernando – and Sarah in particular as my fellow "chicken" person for sharing the various things we had to do. And also Ruth who gave me a valuable introduction to sequence analysis. And also to the other chicken people, Paul, Kieran and Ronan.

I appreciate the help from many others: Karsten for programming solutions, Matteo Fumagalli for sharing data, Olivier Hanotte for giving us a heap of samples and his original donators for getting them in the first place: AK Bhuiyan, Pradeepa Silva, AN Naqvi, Rahamane Sanfo, Racine-Samba Sow and Baitsi Podisi. I would also like to thank Brenda, Sue, Paul and Dave for their time and help over the years – Paul especially for helping with our misguided centrifugation craze.

Various organisations also gave us samples – their generosity was essential for the projects: Vaccine and Infectious Disease Organization (University of Saskatchewan); the Department of Ornithology and Mammalogy, Californian Academy of Sciences (San Francisco); Manor Farms (Co. Monaghan); Donal Campion (Wallslough Farm, Co. Kilkenny); Billy Wilson (Co. Antrim); and Tommy Haran (Co. Meath).

Lastly, I would like to say thank you to my family and friends for their support, encouragement and patience along the long way, particularly to my Mum and Dad, Sally, Killian, Ian – and also for help with proofreading this.

TABLE OF CONTENTS

CHAPTER AND SECTION PAGE NUMBER

Figures

Page

Tables

Page

Abbreviations

Analysis of molecular variance	AMOVA
Average nucleotide differences per site	π
Base pair	Bp
Biased gene conversion	BGC
Coding sequence	CDS
Coding SNP	cSNP
Coefficient of selection	s
Degrees Celsius	^{o}C
Derived allele frequency	DAF
Effective population size	N_e
Interferon-γ	IFNG
Interleukin-1 beta	IL1B
Interleukin-4 receptor alpha-chain gene	IL-4Rα
Jungle fowl	JF
Kilobase	kb
Leucine-rich repeat	LRR
Likelihood ratio test	LRT
Linkage disequilibrium	LD
Megabase	Mb
Million years ago	Mya
Minimum number of recombination events	RM
Nonsynonymous SNP	nsSNP
Pathogen recognition receptor	PRR
Pathogen-associated molecular pattern	PAMP
Polymerase chain reaction	PCR
Recombination rate	R
Salmonella enterica	SE
Single nucletide polymorphism	SNP
Synonymous SNP	sSNP
The relative rate of nonsynonymous to synonymous substitution	ω
Thousands of years ago	kya
Toll/interleukin-1 receptor	TIR
Toll-like receptor	TLR
University of California Santa Cruz	UCSC
Watterson's estimator of genetic diversity	θ_W

CHAPTER 1

Introduction

1.1 Outline and scope of the thesis

This thesis explores diversity among immune genes in chicken at a population and genomic level. Genes associated with host defence determine the result of infectious challenges, so their investigation is likely to identify functionally relevant variation. This approach also permits the evaluation of the chicken's population history and its variability in global populations. A novel combination of intra- and inter-specific tests was implemented here. These analyses enhance and clarify our understanding of the factors defining diversity at immune genes in chicken.

This introductory chapter details the origin of the domestic chicken, its immune genes, and how testing for selection can identify variation of interest. Chapter 2 deals with a survey of polymorphic sites in chicken expressed sequence tags (ESTs). Evidence for selection at immune genes in the avian lineage is analysed in Chapter 3. Chapter 4 describes how two key immune genes were sequenced and that data was analysed in diverse chicken populations and jungle fowl (JF). A genome-wide search for genes under selection and the subsequent sequencing of one immune-related candidate is reported in Chapter 5. Chapter 6 considers the results of sequencing an innate immune locus in global chicken samples, including commercial birds. In Chapter 7, two categories of immune receptor and mediator genes are sequenced using newly developed high-throughput technologies in order to determine the evolutionary patterns at each class. Chapter 8 concludes the thesis by summarising the underlying themes of chicken demographic history and the processes shaping variation at immune genes.

1.2 The chicken and its demographic history

The advent of chicken population genomics began with the publication of the genome sequence of an inbred female red JF (*Gallus gallus*) bird from UCD (University of California-Davis) strain 001 (International Chicken Genome Sequencing Consortium 2004). This bird was originally of Malaysian stock that was bred at a Hawaiian zoo before being developed for genetic and immunological research (Delany 2004). It was also the first livestock organism to be sequenced and is the primary non-mammalian vertebrate model for studying disease (International Chicken Genome Sequencing Consortium 2004).

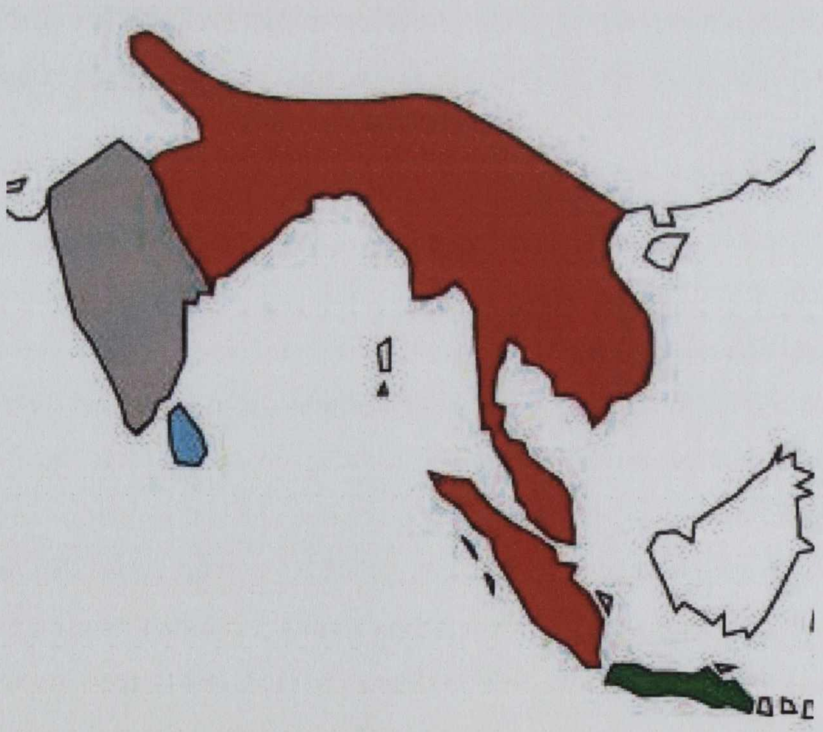
As the first non-mammalian amniote genome to be sequenced, the chicken links the genomics of amphibians, reptiles and fish with that of mammals (International Chicken Genome Sequencing Consortium 2004). The chicken genome has improved our comprehension of vertebrate evolution: for example, in the evolution of amniote sex determination (Smith et al. 2009). Taxonomically, chicken is classified in the Class *Aves* and its diapsid ancestors are estimated to have diverged from their synapsid premammalian common ancestors about 310 mya (million years ago; Kumar & Hedges 1998). Mammals are the most closely related cladistic class to birds: the last common ancestors of amphibians, birds and mammals lived approximately 360 mya (Hedges 2002).

Chickens belong to the subclass *Neornithes*, which contain more than 9,700 extant species (Hedges 2002). The sole other bird species whose genome is sequenced, the zebra finch (*Taeniopygia guttata*; http://songbirdgenome.org), is categorised in class *Passeriformes* (Hackett et al. 2008), and thus shares ancient common ancestry with the chicken, estimated at about 100 mya (Kaiser et al. 2007). Within the *Neornithes*, infraclass *Galloanserae* contains all fowl, including chicken's phylogenetic order, *Galliformes*, as well as other birds such as waterfowl (Chubb 2004). *Galliformes* contains a diverse array of birds, including guineafowl and species of new world quail (Pereira et al. 2002). The genus *Gallus* is one of several bird species groups in the subfamily *Gallininae* of *Phasianidae*, which include old world quail, pheasants, partridges, francolins and peafowl (Kriegs et al. 2007). The genetically most closely related species to *Gallus* fowl whose DNA has been studied are the grey francolin

(*Francolinus pondicerianus interpositus*) and the bamboo partridge (*Bambusicola thoracica*; Kaiser et al. 2005, Kolm et al. 2007).

The jungle fowl genus is composed of four species: grey JF (*Gallus sonneratii*), Ceylon JF (*Gallus lafayetii*), green JF (*Gallus varius*) and the chicken's main wild ancestor, red JF (*Gallus gallus*; Fumihito et al. 1994). Geographically, there is little overlap between these species' ranges: Ceylon JF inhabits the island of Sri Lanka off the coast of India; green JF the island of Java near the south-east Asian continent; grey JF the southern part of the Indian subcontinent; and red JF south and south-east Asia, including islands such as Sumatra (Figure 1.1; Madge et al. 2002). Within red JF, there are geographically defined subspecies: *Gallus gallus* (*G. g.*) *bankiva* on Java island (co-inhabiting sympatrically with green JF), G. g. jabouillei around modern-day Vietnam, *G. g. murghi* on the Indian subcontinent, *G. g. gallus* in south-east Asia, and *G. g. spadiceus* in modern-day Burma (Madge et al. 2002). These subspecies show great variability (Fumihito et al. 1994) and distinguishing between them genetically can be challenging (Kanginakudru et al. 2008).

Figure 1.1 Geographic ranges of red, grey, Ceylon and green jungle fowl.



Ranges for red JF are red; grey JF, grey; Ceylon JF, blue; and green JF, green. This figure was adapted from Figure 1 in Eriksson et al. (2008).

Although red JF was the major donor of chicken genetic diversity (International Chicken Polymorphism Map Consortium 2004), notable contributions have been made by other JF. The clearest example is leg colour: grey JF have yellow legs and red JF have white legs, yet a majority of chickens have yellow legs. Genetic and expression analysis revealed that grey JF and chickens have mutations that stop the action of β-carotene dioxygenase 2, and so possess this leg trait that was historically preferred by many human groups but did not originate in red JF (Eriksson et al. 2008). Analysis of mtDNA segments suggests further interbreeding of grey and red JF, and between grey and Ceylon JF too (Nishibori et al. 2005, Silva et al. 2008). There is evidence that red JF have bred with Ceylon JF as well: phylogenetic trees of the ornithine carbamoyltransferase gene Ceylon, grey and red JF clustered closely to one another as well as chicken (Nishibori et al. 2005).

An initial examination of noncoding and mtDNA suggested that red JF was the main ancestor of chicken (Fumihito et al. 1994, Fumihito et al. 1996). This was later confirmed by more extensive mtDNA analysis revealing the chicken's origins: this analysis indicated that chickens underwent multiple domestication events in south, east and south-east Asia (Liu et al. 2006). Archaeological evidence suggests this occurred in China at least as early as 8 kya (West and Zhou 1989) and perhaps earlier (Nishibori et al. 2005), coinciding with the widespread domestication of many animals and plants after the Younger Dryas era (Salamini et al. 2002). It has not yet been refuted that chickens were domesticated in other locations where red JF are endemic (Fumihito et al. 1996).

In addition to multiple migrations of domestic chickens into other continents and regions far from the domestication centres (West and Zhou 1989), mtDNA evidence suggests there has been considerable historical chicken gene flow across Europe, Asia, Africa and Oceania (Liu et al. 2006). Archaeological evidence indicates domestic chickens reached east Africa at least 3 kya, and that there were three or more independent migrations into west Africa from 1.5 kya (Williamson 2000). Given that red JF are genetically diverse (Fumihito et al. 1994), these movements have meant that modern chicken populations also have elevated variation (Hillel et al. 2003), even though only a small number of key clades may have widely spread since their domestication (Liu et al. 2006). Comparing the complete red JF 1,060 megabase

(Mb) genome to portions sequenced for a broiler (bred for meat), a layer (bred for eggs) and a silkie (bred for mating) revealed 2.8 million variable sites (International Chicken Polymorphism Map Consortium 2004). This comparison showed that diversity between wild and domestic birds is high in comparison to other organisms with over 5 single nucleotide polymorphisms (SNPs) per kilobase (kb). Significantly, this investigation also demonstrated that the chicken did not undergo a major genetic bottleneck during domestication: much of the variation present in extant fowl was present prior to this, and coalesces to 1.4 mya (International Chicken Polymorphism Map Consortium 2004).

As a consequence of the chicken's history of multiple domestications, interbreeding with JF and human-driven migration, its demographic and genetic history may be complex. Evidence for this comes from studies of specific populations where the chicken's genetic variation is constituted by an array of sources, including those from distant continents: for example, Zimbabwean chickens originate in China as well as Africa (Muchadeyi et al. 2008). Domestic chickens, be they Magalasy (Razafindraibe et al. 2009), Chinese (Berthouly et al. 2008, Bao et al. 2008), Indian (Kanginakudru et al. 2008), Chilean or Polynesian (Gongora et al. 2007), Japanese or Korean (Oka et al. 2007), are difficult to genetically distinguish from wild red JF and have multiple separate inputs of variation. However, microsatellite markers may be effective for discrimination within certain populations (Mwacharo et al. 2007).

The chicken is by far the most important bird in terms of both the sheer extent of its worldwide breeding and farming as a food protein resource (McPherson et al.), and the depth of scientific analysis conducted (Zongker 2006). Apart from zebra finch the only birds to be studied in any detail and also extensively farmed are the domestic turkey (*Meleagris gallopavo*) and duck (*Anas platyrhynchos*). The chicken's utility for monitoring infection development *in ovo*, rather than *in utero*, led to its role as a primary model organism for the study of viral and bacterial disease (McPherson et al.). This trait, coupled with its significance as a major food source and a source of zoonotic infections has created a necessity to further study the population dynamics and evolution of genes involved in immune defence in chicken.

1.3 Chicken immune genes and diseases

1.3.1 The immune system and infectious disease:

The study of variation in chicken immune genes is inherently interesting because of their relevance for disease and to illuminate evolutionary history. Their importance is further highlighted by the chicken's potential threat to human health as a reservoir and vector of zoonotic disease (Diamond 2002). Avian illnesses help to generate common pathogens for humans and serve as origins for human diseases, such as the avian influenza virus (Xing et al. 2008) and the severe acute respiratory syndrome coronavirus (SARS). These emerging diseases provide a new impetus to investigate chicken immunity – in particular the relationship between genetic diversity and disease susceptibility, with a view towards developing chickens more resistant to disease (Kaiser et al. 2009). Additionally, many human vaccines are created using chicken cells and eggs (International Chicken Genome Sequencing Consortium 2004).

Although the chicken serves as a model for the study of disease progression, much of the likely function and organisation of its immune system is not yet adequately understood and so is best inferred from that of the human (Burt 2005). Despite existing in identical environments, the immune system of mammals is more complex than that of birds: yet still there are key similarities in patterns of mechanisms found in basal jawless vertebrates (Guo et al. 2009). Avian immunity has two components: an innate one that operates as a fast, generalised response system; and an adaptive part, which fights pathogens in a highly specific manner (Medzhitov & Janeway 2000). Both sides execute their reactions to microbial attack through humoral and cellular responses – the latter activates macrophage, natural killer and cytotoxic T lymphocyte (T_H1) cells, which can cause apoptosis and kill intracellular microbes. Cell-mediated immunity also instigates the expression of signalling molecules, such as interferon-γ and tumour necrosis factor β, in addition to other cytokines that communicate with other cell types (Janeway & Medzhitov 2002). In humoral immunity, T_H2 cells stimulate B cells to produce antibodies and certain cytokines – for chicken, these include interleukins (ILs) 4, 5, 6, 10 and 13 (Kaiser 2007). However, the innate and adaptive components of immunity are not separate; they share signalling molecules and act in synergy to eliminate pathogens (for example, IL10; Mege et al. 2006).

The chicken immune system battles many diseases, including avian leukosis virus, *Brucella abortus*, *Campylobacter jejuni, Clostridum perfringens,* species of *Eimeria, Escherichia coli*, *Haemophilus paragallinarum*, infectious bursal disease virus, *Listeria monocytogenes*, Marek's disease virus, *Mycobacterium avium,* Rous sarcoma virus, different *Salmonella enterica* (*SE*) serotypes, *Staphylococcus aureus* and vesicular stomatitis virus (see Appendix C and Chapter 7 for full lists of references). All of these pathogens alter chicken immune gene expression upon infection and several have host alleles implicated in resistance or susceptibility: Appendix C contains a list of 104 of such genes and the linked diseases, as well as GenBank (www.ncbi.nlm.nih.gov) accession numbers and details of the association. The complex mechanisms of avian host defence operate through the humoral and cell-mediated responses of the innate and adaptive immune systems outlined above and have been explored in detail elsewhere (see Zekarias et al. 2002, Davison 2003).

1.3.2 The evolutionary pressures on the avian immune system:

The geographic distribution, population densities and disease epidemiology of chickens changed dramatically during and since their domestication, undoubtedly shaping their genetic diversity. Novel diseases and increased incidence of infection would have challenged the chicken immune response, necessitating adaptive evolution at key genes (Diamond 2002). Diseases are likely to have played a significant role in the evolution of the chicken genome, especially as a result of domestication, where new challenges would have required adaptation at genes involved in immunity. Changes in population density and distribution, as well as proximity to other species, would have altered the characteristics and scope of diseases affecting chickens.

Immune system genes generally are subject to selective processes – for example in *Drosophila* (Schlenke & Begun 2003, Sackton et al. 2007). Higher diversity at nonsynonymous sites relative to synonymous ones has been observed at immune system loci in mammals when compared to non-immune genes (Hughes et al. 2005), a sign of adaptive evolution (Yang 1997). Human genes implicated in immune defence are more frequently subject to selective processes than the average (Akey et al. 2004, Harris & Meyer 2006): many of these have been subject to recent positive selective sweeps (Williamson et al. 2007) and have been subject to local niche-specific adaptive pressure.

There is evidence of adaptation of chicken immune genes: genes involved in the immune response have lower sequence conservation than other functional categories when compared to the human (International Chicken Genome Sequencing Consortium 2004). Several studies have reported the association of allelic variation at particular chicken immune genes with susceptibility to infection: for example, different alleles at the chicken MHC-B locus are known to alter susceptibility to a diverse array of diseases (Worley et al. 2008). A polymorphism (S631N) in the myxovirus (Mx) resistance protein determines susceptibility to the virus in chicken populations (Ko et al. 2002, Li et al. 2006): this is segregating in indigenous Malagasy (Razafindraibe et al. 2008) and other (Balkissoon et al. 2007) chicken populations as well as red JF (Seyama et al. 2006). There is strong evidence for selection at this gene (Hou et al. 2007, Berlin et al. 2008). Different immune gene variants determine the outcome of infection in chickens (Ye et al. 2006): for example, Tanzanian chickens display variability in resistance to *SE* serovar Gallinarum (Msoffe et al. 2006). However, the evolutionary history of genetic variability in response to avian disease is not fully understood.

It is estimated that 10-20% of human amino acid substitutions are advantageous, about 28% are neutral or nearly neutral, between 30-42% are mildly damaging, and less than 1% are highly deleterious or lethal (Boyko et al. 2008). Similarly, a recent survey of cranially expressed chicken-zebra finch orthologs indicates that about 20% of nonsynonymous changes have been fixed by positive selection during avian history and a further 23% currently segregating in chicken could be mildly deleterious (Axelsson & Ellegren 2009). Consequently, a substantial number of substitutions in immune genes are likely to be adaptive, and given immune genes' higher relative rate of nonsynonymous mutations, they are likely to have proportionately more adaptive polymorphisms that are functionally relevant to chicken immunity. For these reasons, genes involved in the immune system represent appealing candidates for examining the selective processes shaping genetic diversity.

The range of new pathogenic challenges generated by domestication would have necessitated adaptive evolution at chicken genes involved in host defence. These selective forces have shaped variation at chicken immune genes and controlled the chicken's development of a unique repertoire of immunity-related genes (International Chicken Genome Sequencing Consortium 2004, Kaiser et al. 2008, Temperly et al. 2008). Consequently, exploring the evolutionary history of chicken immune genes and searching for evidence of selection can identify genes that are of functional relevance. In addition, understanding how selection operates on specific genes can illustrate how immune genes react to pathogens at a molecular level. Moreover, the isolation of key nucleotide or amino acid sites that determine the effectiveness of the immune response to infection could lead to improved breeding of disease-resistant chicken lines (Kaiser et al. 2009).

1.4 Detecting selection

1.4.1 Defining a neutral model:

Determining the effect of selection on a locus is dependent on the extent and distribution of variation present, the models selected to test the hypothesis, the expected types of the selective processes and the prior knowledge of other forces affecting diversity. The central process is to examine the capacity of given models of evolution in defined demographic contexts to explain the pattern of variability associated with the gene data. By understanding the nature of the allele frequency spectrum, it is possible to develop an expectation of how neutral variation is likely to behave (see Akey 2009).

Many population-based tests for selection evaluate if observed variability deviates significantly from neutrality. Hence, it is crucial to develop models for neutral sets of samples: these are based on assumptions concerning the number and dynamics of alleles likely to be present. At a neutrally evolving locus, it is possible to show firstly that most nucleotides will not vary; secondly that the random effects of neutral drift will yield a predictable distribution of allelic variation; and thirdly that these parameters are computable in the context of different modes of selection (Kimura & Crow 1964). For a population with a neutral stepwise mutation rate, a sample's genotype can vary (or mutate) amongst a defined set of alleles in a linear manner according to the direction of the substitutions (Ohta & Kimura 1973).

For a neutral population, a finite number of alleles would allow it to attain an equilibrium rate of heterozygosity; however an infinite amount of alleles would not (Kimura & Ohta 1976). Different tests implemented here assume either a finite or an infinite allele model: for the former, this is the number of samples resequenced. For the latter, the high allelic diversity of chicken populations reduces this possible bias (International Chicken Polymorphism Map Consortium 2004), and such a model could perform better for non-stepwise mutations invoking recombination or other relevant evolutionary phenomena (Li 1976). Even a very small proportion of neutral mutations that occur in a non-stepwise manner will sharply change the allelic distribution and number (Li 1976). For example, recombination is one method of generating diversity at immune genes like the MHC (O'Neill et al. 2009, Bernatchez

& Landry 2003). Neither finite or infinite sites models fully account for the effect of sampling implemented in this thesis, where a small number of genotypes were sampled from highly diverse populations whose effective population sizes and mutations rates have been estimated but are not necessarily accurately known (Axelsson et al. 2005). However, this uncertainty is a common limitation for analyses based on population-level resequencing and the use of large sample sizes and whole-gene resequencing minimises this effect (Jensen et al. 2007).

The demographic forces outlined earlier clearly have substantial effects on the capacity to detect selection at chicken immune genes, reducing population substructure in particular (Kanginakudru et al. 2008). This is an advantage for modelling adaptation because of the assumption that a neutral population is panmictic largely holds true for chickens. An additional consequence of domestication is that chicken population size is likely to have increased: this growth may mimic signals of directional selection (Akey 2009), and would have increased the abundance of alleles present (Kimura & Ohta 1976).

As a result of the predictability of a neutral allele frequency spectrum, it is possible to determine a range of parameters for which neutrality applies. Additionally, although a subset of mutations may not be completely neutral, in a large population their behaviour may not be different from neutral variation (Crow 1976). It is important to detect functional variation that may enhance or compromise resistance to disease: substitutions of interest are those which are either deleterious or advantageous such that the strength of selection is significantly different from that of neutral diversity. Alleles that are severely deleterious are lethal and are unlikely to manifest in the populations; equally, those that are very positively selected sweep quickly to fixation and are difficult to detect at an intraspecific level unless very recent (Akey 2009). Therefore, the strategies to detect adaptation here combine the assessment of intraspecific variation, where changes in genetic fitness are likely to be discrete and functionally relevant to extant birds (for example, Berlin et al. 2008), with interspecies tests that can identify significant historical changes that have swept to fixation in the avian lineage (Yang 2002).

1.4.2 Patterns of adaptation:

Understanding variation at a locus is dependent not only on the neutral model but also on the alternate models, which are related to the type of selection expected to be present. In a population whose effective size is N_e, the effect of a mutation can be quantified as a selection coefficient value (s) where $-1 \leq s \leq 1$ such that a neutral effect has $s = 0$. Advantageous polymorphisms have $s > 0$ and deleterious ones $s < 0$, however, if $-1/(2N_e) \leq s \leq 1/(2N_e)$, then the selective pressure caused by the variant is effectively neutral (Kimura 1979). For chicken, which is likely to have a very large N_e, this means the effects of a significant proportion of genome-wide changes are no different from neutral ones (Hurst 2009, Eyre-Walker & Keightley 2007). Substitutions with negative consequences such that $s < -1/(2N_e)$ are likely to be eliminated from the population: this is termed purifying or background selection (Hughes 2007). Similarly, for advantageous variants with $s > 1/(2N_e)$ are directionally selected, increasing the allele frequency and ultimately leading to fixation (in an ideal population where all other variation is neutral; Akey et al. 2004). A corollary of nearly neutral theory is that mildly deleterious mutations will not always be eliminated and consequently will segregate in populations (Ellegren 2008). Likewise, many nearly neutral mutations will be fixed by genetic drift (Conant 2009).

The dynamics of certain polymorphisms are more nuanced: particularly at immune genes where mutations induced in host defences by the generation of new pathogen challenges in changing local environments may vary not solely according to the infection pathogenicity, but also according to the gene's role and functional constraints (Sackton et al. 2007). Compounding the genetic consequences of these selective forces are the number and variety of microorganisms to which the chicken must respond concurrently, and the introduction of immigrant red JF genotypes into the populations (Kanginakudru et al. 2008). Therefore, complex models of evolution may fit the evolution of these genes more exactly: these models invoke selection that is dependent on the relative frequencies of different alleles (Charlesworth 2006). Balanced diversity can also emerge from the latent admixture of previously separate groups, and in situations where a heterozygous allele confers more fitness than either homozygote combination (Charlesworth 2006).

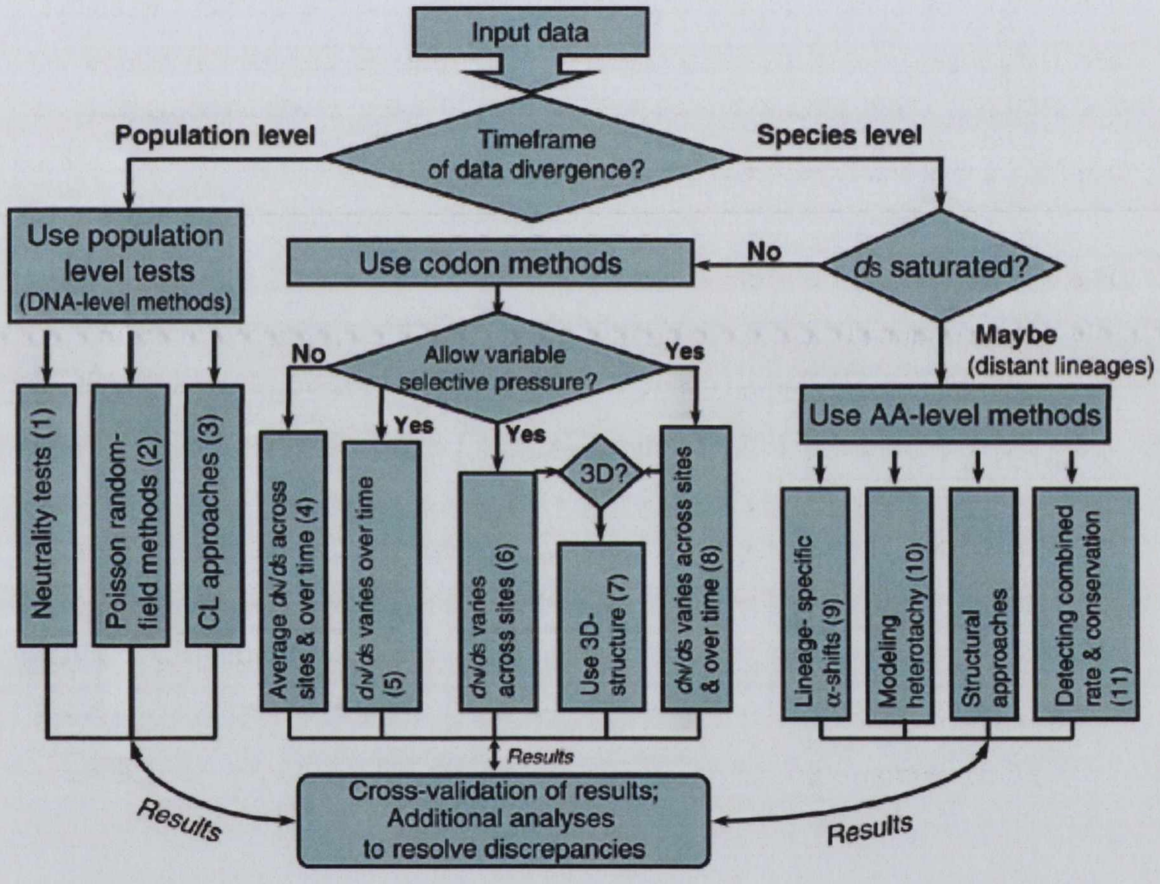
Balancing selection also operates on variants that were previously neutral or have been historically selected in a variegated manner (Innan & Kim 2004). The consequences of fluctuating selection on diverse sets of alleles in a population differ considerably to a simplistic positive selection model of a single pathogen and resistant allele (Przeworski et al. 2005). Classically, a single new mutant allele is favoured; however, if the selected variant has more than one copy in the population, or is present in a set of migrant genotypes, the subsequent trend of variation is different (Hermisson & Pennings 2005). Selection acting on standing diversity that was previously neutral will cause much less variation to be lost than would be expected for a selective sweep of a sole *de novo* allele (Pennings & Hermisson 2006a). As a result, the effects of directional selection will be weaker: in the context of serial selective sweeps on immune genes in a dynamic, pathogen-rich environment this would mean less diversity is lost and that the magnitude of the selection signature would be considerably reduced (Innan & Kim 2004).

Analyses of other disease-associated chicken genes have identified a pattern of balancing selection at the chicken MHC (Worley et al. 2008). This is a common pattern of variation at vertebrate MHC genes and could be caused by the wide variety of infections to which the MHC responds (Jeffery & Bangham 2000). This is a form of frequency-dependent selection (Asthana et al. 2005), where MHC variants that are highly resistant to disease rise in frequency in the population while pathogens evolve to enhanced forms that identify susceptibilities in the host group: if the resistant genotype attains fixation, pathogens will either die out or adapt. Consequently, a form of predator-prey relationship continues, and the populations that preserve diversity at the MHC appear to persist: this may be a general pattern for certain types of immune genes.

1.4.3 Methods to detect selection:

While it is possible to categorise the effects of mutations, the power to detect their selection signatures has been enhanced by the development of many different complementary population-level and genomic approaches (Zeng et al. 2006, Zhai et al. 2008, Ellegren 2009). Primarily, these are based on examining nucleotide changes within a population or between species. Differentiating these into groups of tests can be useful, though where possible in this thesis the results of disparate methods that deal with the same question were analysed together to give a more comprehensive picture of variability. Three such classes for assessing intraspecies diversity outlined here are tests based on firstly linkage disequilibrium (LD), secondly allelic population differentiation, and thirdly the site frequency distribution (Zeng et al. 2007, Hurst 2009). Other approaches explained later include phylogenetic trees and networks, as well as non-parametric sampling methods. A more comprehensive perspective of intra- and inter-species and comparative genomic approaches is available elsewhere (Figure 1.2 – from Figure 1 in Anisimova & Liberles 2007).

Figure 1.2. Schematic of methods to detect selection.


Input data
Population level
Timeframe of data divergence?
Species level
Use population level tests (DNA-level methods)
Use codon methods
No
ds saturated?
No
Allow variable selective pressure?
Yes
Maybe (distant lineages)
Yes
Yes
Use AA-level methods
Neutrality tests (1)
Poisson random-field methods (2)
CL approaches (3)
Average dN/dS across sites & over time (4)
dN/dS varies over time (5)
dN/dS varies across sites (6)
3D?
Use 3D-structure (7)
dN/dS varies across sites & over time (8)
Lineage- specific α-shifts (9)
Modeling heterotachy (10)
Structural approaches
Detecting combined rate & conservation (11)
Results
Results
Cross-validation of results; Additional analyses to resolve discrepancies
Results

This figure is from Figure 1 in Anisimova & Liberles (2007) – for details of the parethesised numbers see this reference. Although not all approaches are implemented in this study, it does illustrate the usefulness of multiple synergistic approaches to identify properties of interest.

LD is the non-random association of alleles: if a pair of genes is linked on the same chromosome of DNA, it is more likely that they will segregate together. This can be disrupted by recombination events during meiosis that mix parental alleles to generate new allele combinations. Thus as the number of generations increases, the probability of a defined pair of alleles associating on the same chromosome decreases assuming both are neutral: if this temporal decay of LD deviates significantly from that expected, it is evidence for selection (Gu et al. 2008). LD-based tests on human genes have proved useful approaches for detecting selection (for example, Akey et al. 2003) but only if the candidate region is well known (Zeng et al. 2006). For chicken immune genes, certain sites or exon domains may be subject to selection when the rest of the gene may not, such as the Mx gene (Berlin et al. 2008, Hou et al. 2007). Additionally, many chicken genes display high levels of population diversity driven by demographic effects, which can exaggerate the rate of recombination, concealing the extent of true LD (Price et al. 2008).

Tests based on nucleotide or allele diversity examine the levels of heterozygosity present and are effective at detecting demographic structure as well as selection (Zeng et al. 2007). For alleles at a single locus, if the relative proportions of homozygotes and heterozygotes in a population are significantly different from those expected, it can be indicative of non-neutral evolution (Hurst 2009). For example, the Ewens-Watterson test examines the levels of heterozygosity in populations in order to determine if they are significantly different from those expected for a given allele distribution (Ewens 1972, Watterson 1978). An excess of heterozygosity may indicate diversifying selection; elevated homozygosity can be symptomatic of positive selection, where an allele has been fixed in the population and so has reduced local diversity. Tests based on the incidence of segregating sites can examine the extent of variation between populations by determining the associated heterozygosity levels: if the populations are very different, it can be a sign of adaptive evolution (Wright 1951). Wright's F-statistics measure the extent of differentiation between groups and can also be used to determine inbreeding (Charpentier et al. 2007), gene flow, population structure and migration rates (Eldon & Wakeley 2009).

A wealth of tests have been developed that incorporate the spectrum of SNP allele frequencies into tests, which arguably are more incisive strategies for examining variation in species exhibiting high diversity with little population structure, such as chicken. These tests included Tajima's *D* (Tajima 1989), Fu's F_s (Fu 1997), Fu and Li's *D* and *F* (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000); the latter three metrics require outgroup information to assign ancestry. These tests examine the relative ratios of singleton, low, medium and high frequency alleles in the population. However, each test incorporates only certain components of the allele frequency spectrum, and so it is most instructive to use them in a complementary fashion (Zeng et al. 2006). For example, combining Tajima's *D*, Fay and Wu's *H* and the Ewens-Watterson test is more efficient than each test alone at detecting positive selection, particularly in the presence of recombination (Zeng et al. 2007). Multiple tests of neutrality were used in concert in this thesis in order to give a more comprehensive description of the pattern of variation in the data. Combining summary statistics can be a more powerful approach (Innan 2006), and so using approaches measuring both the absolute level of polymorphisms and their relative frequency distributions were implemented where useful.

Interspecies tests for selection have been developed: the most basic of these examines the differences between pairs of gene sequences at nonsynonymous and synonymous sites (Yang 1997). Synonymous changes have no protein-level effect and so evolve in a neutral, stochastic manner, whereas nonsynonymous mutations result in amino acid alterations; thus calibrating the rate of change at nonsynonymous sites by that at synonymous sites provides a measure of the rate of adaptive evolution. Although the assumption that synonymous (or noncoding) sites are neutral is not always true (Shields et al. 1988, Pagani et al. 2005), it is generally a valid assumption if N_e is small (Chamary et al. 2006). By conferring phylogenetic relationships onto such a simple model, differing evolutionary rates between lineages can be deduced by determining the relative probabilities of a neutral model and a non-neutral model (Yang 2002). Similarly, sites that could have mutated in a non-neutral manner can be inferred (Anisimova et al. 2001, Yang et al. 2005). These tests can be combined further so that sites positively selected on specific branches can be identified (Yang & Nielsen 2002), though their high detection power may be compromised by a high false positive rate for datasets with low rates of polymorphism (Nozawa et al. 2009). Therefore, genes that have been historically under selection between species can be discovered and these are likely to hold functional relevance in extant birds.

Inter- and intra-specific investigations can be amalgamated so that for a given species, the relative evolutionary rates can be compared with that of a second species.

For example, the McDonald-Kreitman test evaluates the ratio of substitutions in two classes of intercalated sites by comparing the rate between a pair of species to that within one of the species' population (McDonald & Kreitman 1991). If the ratio is significantly higher between species, it is evidence that one class of sites has preferentially been fixed more frequently and thus have been advantageous. This test was originally developed from the Hudson-Kreitman-Aguadé (HKA) test (Hudson et al. 1987) and is commonly implemented for nonsynonymous and synonymous or nonsynonymous and silent polymorphisms (Smith & Eyre-Walker 2002). Such tests can also discriminate between positive selection and the relaxation of selective constraint (Eyre-Walker 2002, Cruz et al. 2008).

Certain genes that determine susceptibility to infection and function in both the innate and adaptive immune responses have been subject to selective forces in the chicken, such as Mx (Ko et al. 2002, Li et al. 2006, Seyama et al. 2006, Hou et al. 2007, Berlin et al. 2008), MHC-B (Worley et al. 2008), and those discussed here later: IL1B (Downing et al. 2009a), IL4RA (Downing et al. 2009b) and lysozyme (Downing et al. 2009c). Pathogen-driven selective pressures have resulted in the expansion and diversification of both leukocyte receptor (Laun et al. 2006), immunoglobulin, and immunoglobulin receptor families in chicken (International Chicken Genome Sequencing Consortium 2004). This indicates that analysing immune genes for evidence of selection in order to evaluate population history and functional variation can enhance our understanding of host defence and the evolutionary mechanisms implemented to combat pathogen virulence. The identification of alleles implicated in diseases could lead to the breeding of chicken flocks resistant to disease.

1.5 Evaluating variation at chicken immune genes

By determining the dynamics of the selective processes acting on chicken immune genes, it is possible to test hypotheses regarding their functional importance, the patterns of variation present and the demographic history of the chicken (International Chicken Genome Sequencing Consortium 2004). It is clear that the intricate demographic history of the chicken must be explored more thoroughly in order to accurately discern the nature of evolutionary pressures acting on immune genes. To this end, a global set of chicken samples, including commercial lines, are studied in this thesis, and tests are incorporated on their population structure and how this is reflected in their diversity. Additionally, variation in red, grey, Ceylon and green JF and their relationships with chicken are examined. Although variability at immune genes has been researched in inbred lines (for example, Ko et al. 2002), functional variation in global chicken populations remains largely unexplored, and thus is a novel attribute of this thesis.

The publication of the chicken genome sequence permits the implementation of genomic approaches to the identification of genes evolving under selection, which may reflect their functional importance. Such approaches have proved illuminating at other disease-associated gene regions like the avian MHC (Kulski & Inoko 2004) and in other organisms (Ronald & Akey 2005). The continued sequencing of other avian species also now allows interspecies investigations of large sets of bird genes, a strategy that may prove more informative than comparisons of birds with mammals, which for many genes display such extensive protein-level sequence divergence that it may conceal relevant functional changes (International Chicken Genome Sequencing Consortium 2004). Similar methods have been successfully performed on other vertebrates but not in birds, so this is an additional feature of this work (for example: Schlenke & Begun 2003, Hughes et al. 2005).

Genes historically under selection between species are likely to remain of significance in modern birds (Smith & Eyre-Walker 2002), and the diversity of such genes in populations was evaluated to determine if they are still evolving in an adaptive manner. These genomic schemes to search for important immune genes are accompanied by a literature survey of chicken genes known to influence resistance and susceptibility to disease. Although the catalogue of avian immune genes is increasingly well documented (Kaiser 2007), a gene- rather than a disease- focused perspective on immunological results can identify pivotal genes, so this is a fresh perspective on the genetic effects of chicken diseases.

The recent development of innovative sequencing technologies has meant that computational genomic analyses can be fused with population-level assessments of chicken diversity based on gene groups rather than single genes. Consequently, patterns among functional classes of immune genes can be elucidated. A population genomics approach to determining the evolutionary mechanisms of key immune gene classes has only been implemented in *Drosophila* (Sackton et al. 2007) and at the vertebrate MHC (Hughes & Yeager 1998). This has been done in a limited form for humans (Ferrer-Admetlla et al. 2008, Fumagalli et al. 2009), but the persistence of a gene-centric rather than a gene class-oriented approach limits our ability to understand more fully selective pressures that operate on different components of the immune system. Therefore, this is an innovative method to understand the underlying patterns in categories of immune genes.

1.5.1 Explaining chicken immune gene history:

Genomic approaches to detecting selection at genes require a comprehensive controlling for and understanding of the consequences of demographic effects in order to minimise the detection of false positives (Hermisson 2009). Consequently, accounting for the diversity at immune genes in chicken must be framed within this context. Commercial broilers were included in the analysis to compare their diversity to that of village chickens. It is possible that human-driven breeding, differential exposure to infectious diseases and selective pressure due to vaccination regimes may have significantly changed their variability.

The genomic and literature-based strategies used in this thesis endeavoured to identify key immune genes. These were resequenced in global populations in order to both to test for selection and to understand chicken population dynamics using multiple intraspecific and interspecies analyses to isolate the properties most relevant to immune defence. These results were consistent with the expectation that avian immune genes have historically been subject to selection, and that the complex demographic history of chicken and JF explains much of the chicken's elevated genetic diversity. These projects culminated in a survey of population variation in immune receptor and cytokine genes that evaluated the evolutionary constraints on these groups in light of their differing roles. This evolutionary immunomic study illustrated that the selective processes acting on vertebrate immune genes are determined by multiple competing effects.

CHAPTER 2

Evidence of the adaptive evolution of immune genes in chicken

2.1 Introduction

Understanding the evolutionary patterns of variability in gene functional categories can illuminate their characteristics. Immune system genes in particular are subject to acute selective pressures in order to resist pathogenic attacks and consequently undergo many protein-level sequence changes. It is known that chicken host defence genes evolve under stronger positive selection than other functional categories of genes: in alignments with human genes, they possess lower sequence conservation (International Chicken Genome Sequencing Consortium 2004). In mammals and insects, genes implicated in immunity have higher diversity at nonsynonymous relative to synonymous sites (Schlenke & Begun 2003, Hughes et al. 2005). In humans, genes associated with defence have a higher fraction of genes subject to positive selection than average, and genes with high rates of nonsynonymous mutations are more frequently associated with disease (Bustamante et al. 2005).

The basis for understanding the characteristics of gene functional categories in chicken has been enhanced by the ongoing sequencing of the zebra finch genome, the second bird species to be extensively sequenced. This sequence provides an avian context for examining how variation in chicken has evolved since its ancestors' divergence from zebra finch as well as well as a calibrating point for studying intraspecific diversity within chicken. Additionally, the lower sequence divergence of the chicken with the zebra finch compared to that with mammalian genomes permits a more precise analysis of functional diversity (Ellegren 2007). As a result, exploring the evolutionary history of chicken immune genes within the avian lineage is more likely to inform on molecular traits that distinguish them from other genes.

Higher GC content in the chicken genome is associated with smaller chromosome sizes (Andreozzi et al. 2001) and higher rates of nucleotide substitution (Webster et al. 2006). GC content in the chicken genome is elevated in regions that are gene dense, a trait shared with mammalian genomes (Constantini et al. 2007). However, the evolution of the chicken genome is unusual because it has been subject to more complex pressures, such as a metabolic incentive to dramatically reduce genome size (Organ et al. 2007, Hughes & Hughes 1995). Consequently, avian genomes could be subject to selective processes to optimise their sizes, chromosome structures and gene distributions.

Immune genes have been subject to many selective processes during their evolutionary history and this gene class was investigated here in a set of orthologous chicken and zebra finch genes with functions assigned from the human ortholog. Tests demonstrated that nonsynonymous sites at immune genes were highly conserved both in chicken and on the avian lineage. McDonald-Kreitman tests provided evidence of adaptive evolution and a higher rate of selection on replacement substitutions at immune genes compared to that at non-immune genes. Further analyses showed that GC content was significantly higher in chicken than in zebra finch genes, and was significantly elevated in both species' immune genes. Pathogen challenges are likely to have driven the selective forces that have shaped variation at chicken immune genes, and continue to restrict diversity in this functional class.

2.2 Methods

2.2.1 Identifying a set of annotated bird genes:

In order to determine a set of functionally annotated chicken genes, translations of chicken gene transcripts downloaded from Ensembl (18,766) (www.ensembl.org/Gallus_gallus/) were searched against 38,754 human protein RefSeqs (www.ncbi.nlm.nih.gov/RefSeq) using a basic local alignment search tool (Blastp; Altschul et al. 1990) to identify single best hit pairs (15,754). These best hits were used as a reference to assign human gene function and process categories from 33,905 Panther human gene entries (Thomas et al. 2003) successfully to 9,910 chicken orthologs.

Zebra finch ESTs and mRNAs (67,671) from GenBank were cleaned of vector contaminants using SeqClean (http://www.tigr.org/tdb/tgi/software/) and repetitive sequences were masked using RepeatMasker (http://www.repeatmasker.org). In this thesis, default parameters were used except where stated. The TIGR gene indices clustering tools (Tgicl; Pertea et al. 2003) clustered zebra finch sequences whose length $\geq$ 100 bases and identity $\geq$ 96% for overlapping regions into 9,716 zebra finch contigs.

Orthologous chicken-zebra finch sequence pairs were identified by searching the zebra finch contigs against the chicken protein gene transcripts using Blastx (Altschul et al. 1990), with an E value $\leq$ e-10 separating best hits for each protein from paralogous sequences. These best-hit protein pairs were aligned with T-Coffee (Notredame et al. 2000) using Perl scripts (ckzfNEWblastx.pl, BLASTX.pl and hitParserZF.pl in Appendix A). Those with length < 70 amino acids or sequence identity < 60% were discarded. This chicken-zebra finch single best hit ortholog data was first published by Downing et al. (2009b), see Chapter 5.

These protein alignments were used as templates to generate 3,653 chicken-zebra finch pairwise coding sequence (CDS) alignments that were cross-referenced with the 9,910 chicken genes with orthologous Panther functions to generate 2,604 annotated chicken-zebra finch gene pairs. 64 of these could be identified confidently as those whose human ortholog had a function or process related to immunity; this was done by examination of the Panther orthologous human gene functional categories and processes. Genes with positions not yet allocated to a defined position on a chromosome were excluded. Only autosomal chromosomes with known chromosome sizes (Gao & Zhang 2006) were considered. The Z and W chromosomes have divergent properties: for example, W is gene poor and Z is gene rich (Marshall Graves 2009, Smith et al. 2009). Their unique evolutionary history as sex chromosomes may affect the dynamics of immune genes located there (International Chicken Genome Sequencing Consortium 2004).

2.2.2 Determining interspecies and intraspecific variation:

Pairwise ratio d_N/d_S (ω) was calculated for each CDS alignment using the codeml implementation of the PAML 3.15 package (Yang et al. 2002) where d_N was the number of nonsynonymous mutations per nonsynonymous site and d_S the number of synonymous substitutions per synonymous site. If synonymous and nonsynonymous mutations are neutral, the relative rates of each are expected to be equal so that $\omega = 1$ (Yang et al. 2002). Departures from this, where $\omega > 1$ ($d_N > d_S$) suggest that nonsynonymous mutations are advantageous, and are maintained under directional selection. If $\omega < 1$ ($d_N < d_S$) then the nonsynonymous SNPs may be deleterious since they are not preserved and are likely to be subject to purifying selection (Yang et al. 2002). GC content at 3rd codon position (GC3) was calculated for each sequence from these alignments: the 1st and 2nd positions are subject to greater purifying selection, so GC3 was a more neutral measure.

Intraspecific rates of evolutionary change were also calculated for the 2,604 functionally annotated chicken genes as P_N/P_S, the ratio of nonsynonymous mutations (P_N; which change the amino acid in the protein sequence) to synonymous mutations (P_S; which cause no amino acid change) per effective CDS coding site (calculated as the CDS length corrected for the coverage divided by the gene length). After adjusting for genome sequencing coverage rates, SNP frequencies and GC3 for genes and immune genes were explored using one-tailed Student's t-tests and using Pearson's correlation coefficient (r), a measure of the shared linear variation between parameters.

2.2.3 McDonald-Kreitman tests for selection:

The McDonald-Kreitman tests (McDonald & Kreitman 1991) were implemented with DnaSP to examine the rates of evolution within a species (chicken here) to that between species (between chicken and zebra finch) at two categories of sites. The relative ratios of fixed nonsynonymous (D_N) and synonymous (D_S) substitutions and polymorphic nonsynonymous (P_N) and synonymous (P_S) changes are evaluated as D_N/D_S and P_N/P_S. Sites are determined to be fixed if they are variable between the species but not within chicken. Polymorphic sites are those that vary solely within the species being tested, thus the test compares rates of interspecies and intraspecific diversity. Nonsynonymous and synonymous sites are intercalated in coding sequences and thus closely follow each other's genealogical history, so the absolute numbers of polymorphisms (D_N/D_S) can be used instead of the rates (d_N/d_S); this also makes the test more robust to recombination. The test calibrates the rates of nonsynonymous site change for what is assumed to be a neutral rate at synonymous sites. Although mutations at nonsynonymous sites can be neutral, deleterious or advantageous, only those that are in the latter category are expected to be preferentially retained.

If $D_N/D_S > P_N/P_S$ or $D_N/D_S < P_N/P_S$ for a significant one-tailed Fisher's Exact Test, it is indicative of non-neutral adaptation (McDonald & Kreitman 1991). Such observations show that there are significant differences in the rates of evolution with species or between species. Classically, if $D_N/D_S > P_N/P_S$, it suggests the presence of adaptive evolution, whereas if $D_N/D_S < P_N/P_S$, it is more consistent with background selection on the ancestral interspecies branch (Eyre-Walker 2002).

An observed fixation index (*FI*) for all genes and subsets was also determined as:

$$FI = \frac{(D_N / D_S)}{(P_N / P_S)}$$

reflecting the McDonald Kreitman test. If neutral, *FI* should approximate a value of 1; however, this may be violated in regions of relaxed selective constraint (Smith & Eyre-Walker 2002). Consequently, the expected contingency table values of D_N, D_S, P_N and P_S for each gene were determined and summed across all genes so that an expected fixation index (*eFI*) could be calculated as outlined by Axelsson & Ellegren (2009). This also allows an estimation of the fraction of nonsynonymous mutations driven by positive selection (α) to fixation as:

$$\alpha = \frac{(FI - eFI)}{eFI}$$

such that *eFI* represents an unbiased estimate of the neutral evolutionary rate against which *FI* can be compared.

On the basis that $P_N = 4N_e\mu fL_Nk$ and $P_S = 4N_e\mu L_Sk$, where L_S was the total number of synonymous sites, L_N was the total number of nonsynonymous sites, μ is the mutation rate per base, and k is a constant dependent the chances of observing a neutral allele (see Smith & Eyre-Walker 2002), the mean proportion of amino acid-altering neutral substitutions was determined as

$$f = P_NL_S/P_SL_N$$

This did, however, assume that f was relatively constant since the chicken-zebra finch divergence time.

2.3 Results

Protein and coding sequences for a set of chicken genes whose functions were determined from human orthologs were aligned to 2,604 orthologous zebra finch contigs clustered from EST and mRNA sequences. A series of interspecies and intraspecific analyses were conducted in order to test for evidence of selection in chicken immune genes.

2.3.1 Conservation at chicken immune genes:

The analysis included 410,735 SNPs distributed across the autosomal chicken genome at a rate of 0.011 per kb of transcript covered, a number lower than reported elsewhere (International Chicken Polymorphism Map Consortium 2004) because only chicken genes with both zebra finch and functionally annotated human orthologs were investigated. 8,848 of these SNPs were in immune genes: 17 of these were nonsynonymous and 129 were synonymous. In comparison 1,276 nonsynonymous and 4,940 synonymous SNPs were identified in 401,728 SNPs at non-immune genes.

Comparisons of diversity between groups within chicken showed that the average P_N/P_S (mean 0.059 vs 0.121 for non-immune; Table 2.1) and number of nonsynonymous substitutions per gene (0.266 vs 0.503 for non-immune) were both about twice as high for immune genes as they were for non-immune genes. Although the rate of synonymous substitutions per gene was about the same for each group (2.016 vs 1.989 for non-immune), the rate of fixation of neutral amino acid-changing variants was much lower in immune than in non-immune genes ($f = 0.021$ vs 0.050 for non-immune; Table 2.1). These results illustrated that nonsynonymous sites within chicken were more conserved at immune genes.

Alignments of chicken and zebra finch genes found that the average ω value (0.096) was about the same as that observed between a red jungle fowl and a broiler for genomic mRNA transcripts (0.098; International Chicken Polymorphism Map Consortium 2004), and in an analysis of cranially expressed chicken-zebra finch gene pairs (0.085; Axelsson et al. 2008), suggesting that the gene dataset was not biased (Axelsson & Ellegren 2009). Mean ω values were higher for non-immune (0.097;

Table 2.1) than immune (0.083) genes, signifying conservation of nonsynonymous sites in the avian lineage at immune genes as well.

Table 2.1. Mean intra- and inter-specific diversity for chicken and zebra finch at all, immune, non-immune and McDonald-Kreitman test outlier genes.

Gene set	All	Immune	Non-immune	Genes with p < 0.05 [1]
Number	2,604	64	2540	26
ω [2]	0.0963 ± 0.130	0.0826 ± 0.091	0.0967 ± 0.131	0.2950 ± 0.169
Chicken GC3	0.600 ± 0.173	0.652 ± 0.171	0.599 ± 0.173	0.518 ± 0.132
Zebra finch GC3	0.554 ± 0.159	0.608 ± 0.182	0.553 ± 0.158	0.507 ± 0.133
D_N	94,635	1,504	93,131	1096
P_N	1,293	17	1,276	0
D_S	384,749	5,852	378,897	1439
P_S	5,069	129	4,940	272
P_N per kb [3]	0.459	0.327	0.464	0
P_S per kb [3]	1.813	2.474	1.800	0.111
P_N/P_S	0.255	0.132	0.258	0
D_N/D_S	0.246	0.257	0.246	0.762
L_N/L_S [4]	3.075	3.363	3.068	2.860
FI [5]	0.964	1.950	0.952	0
eFI [6]	1.056	0.945	1.060	1.377
α [7]	-0.087	1.062	-0.102	-1.000
Coverage [8]	0.814	0.723	0.816	0.859

[1] Genes whose McDonald-Kreitman test one tailed p values < 0.05 for $D_N/D_S > P_N/P_S$. [2] Calculation excluded non-immune gene XM_422655 that had $d_N > 0$ and $d_S = 0$. [3] Per kb of effective CDS nucleotide length. [4] Total number of synonymous (L_S) nonsynonymous (L_N) sites. [5] Observed fixation index, $FI = (D_N/D_S)/(P_N/P_S)$. [6] Expected fixation index, eFI. [7] Proportion of fixed nonsynonymous mutations driven by positive selection fixed in chicken, $\alpha = (FI - eFI)/eFI$. [8] Mean transcript coverage per base.

2.3.2 Adaptive evolution in the chicken lineage:

Genes that had a higher ratio of fixed nonsynonymous to synonymous substitutions (D_N/D_S) compared to the ratio of segregating nonsynonymous to synonymous substitutions (P_N/P_S) may be have undergone adaptive evolution (McDonald-Kreitman 1991). McDonald-Kreitman tests on the set of immune genes showed a significant excess of fixed nonsynonymous changes on the chicken-zebra finch lineage (*FI = 1.95*; one-tailed p = 0.004) that was not present for non-immune genes, whose *FI* value was about two times lower (*0.95)*. D_N/D_S for non-immune (0.246; Table 2.1) and immune (0.257) genes were about equal, but P_N/P_S was much higher for non-immune genes (0.258 vs 0.132 for immune genes). The high number of SNPs per immune gene ensured that this largely unlinked set of loci should be robust to aggregative McDonald-Kreitman tests (Schlenke & Begun 2003, Andolfatto 2008). The mean fraction of neutral amino-acid replacement mutations (*f*) for each gene with $P_N > 0$ and $P_S > 0$ was not different between those with immune (0.222) and non-immune (0.239) functions.

An unbiased estimate of the neutral rate of the fixation of amino acid changing variants in chicken, *eFI*, was lower for immune (0.95) than non-immune (1.06) genes, further illustrating that immune genes were more conserved than non-immune ones. *eFI* for all genes (1.06) was of the same scale as other datasets (Axelsson & Ellegren 2009). Given the immune set's much higher *FI*, the estimated proportion of amino acid changes fixed in chicken that were driven by positive selection ($\alpha = (FI - eFI)/eFI$) was much higher for immune (1.06) than non-immune genes (-0.10). This indicated that immune genes were subject to stronger selective processes and also that there were deleterious alleles present at non-immune genes.

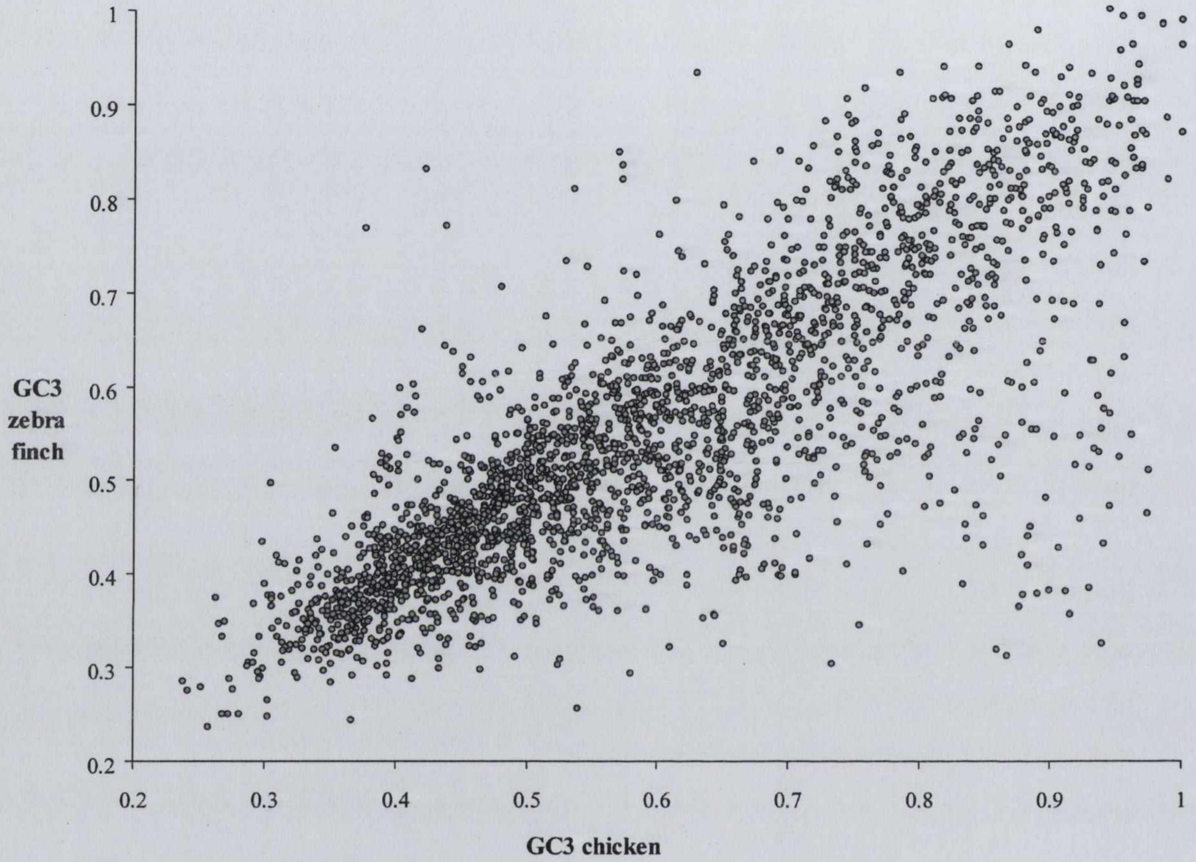
Individual McDonald-Kreitman tests on chicken genes identified 26 (1% of the total) with a significantly higher D_N/D_S than P_N/P_S (one-tailed $p < 0.05$). This set of genes had an average coverage rate (0.86; Table 2.1) above that for all genes (0.81), indicating that the absence of the detection of nonsynonymous SNPs segregating in chicken was not due to poor coverage. Although this group had an average ω significantly higher than that for all genes (mean 0.295 vs 0.096 for all, $p < 1 \times 10^{-6}$; Table 2.1), no replacement mutations were found segregating in the chicken population, suggesting that the significant McDonald-Kreitman tests may be detecting severe purifying selection rather than adaptive evolution. This group contained an immunity-related helicase (KU70, McDonald-Kreitman $p = 0.021$) and a DNA polymerase (eta, McDonald-Kreitman $p = 1.8 \times 10^{-5}$) involved in homologous recombination during DNA repair (Faure et al. 2008) and synthesis (Kawamoto et al. 2005), respectively.

2.3.3 GC content higher in immune genes:

GC3 was significantly higher for immune genes than for non-immune genes in both chicken (mean 0.65 vs 0.60 for non-immune, $p = 0.016$) and zebra finch (mean 0.61

vs 0.55 for non-immune, p = 0.006). GC3 was significantly higher for chicken than zebra finch genes (0.60 vs 0.55, $p < 1 \times 10^{-6}$; Table 2.1), though it was highly correlated between the species, as expected ($r^2 = 0.940$, $p < 1 \times 10^{-6}$; Figure 2.1). Gene rates for GC3 and ω did not correlate significantly.

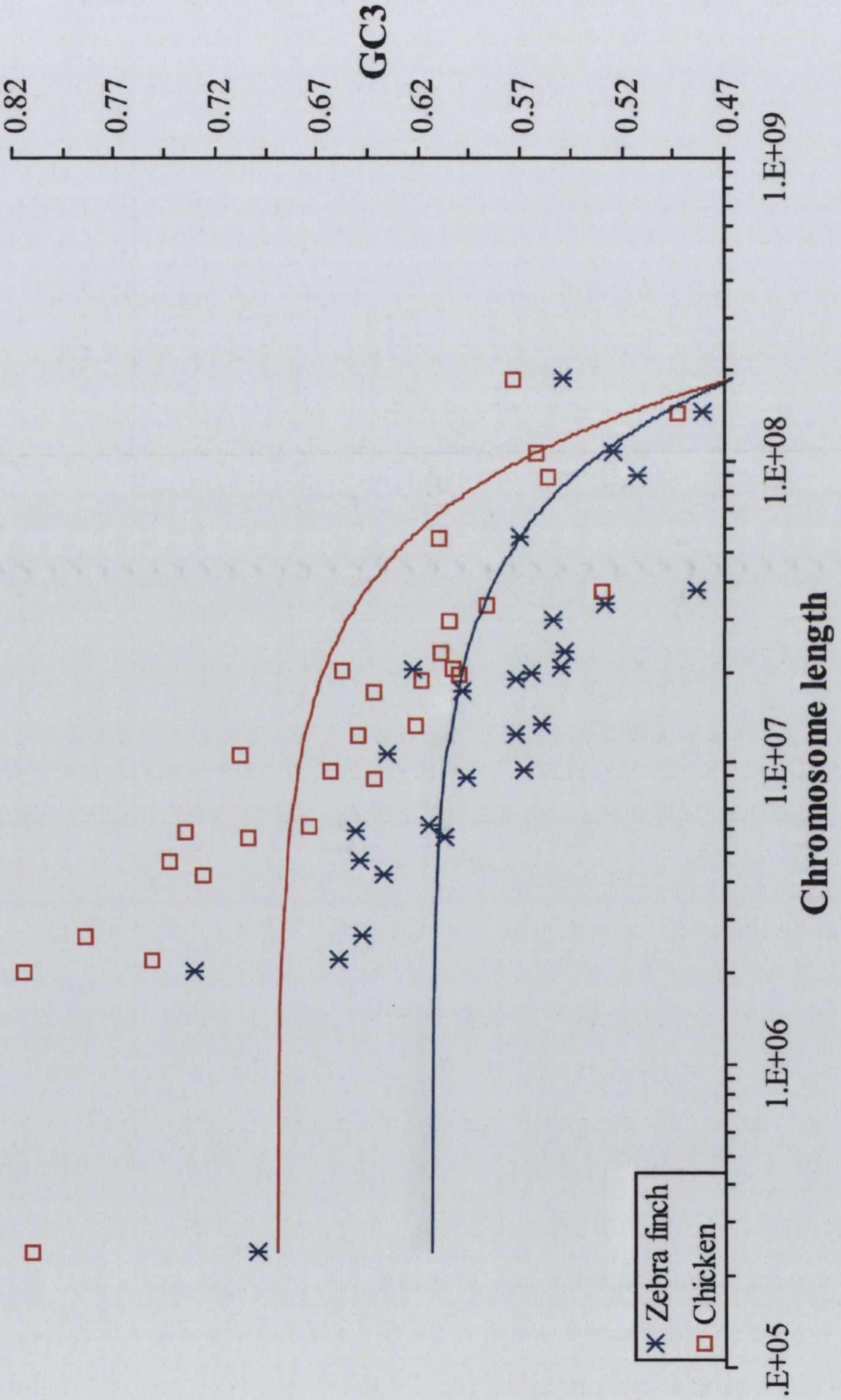
Figure 2.1. Correlation of GC3 content at chicken and zebra finch genes.



The best fitting linear correlation (not shown) has $r^2 = 0.940$ ($p < 1 \times 10^{-6}$).

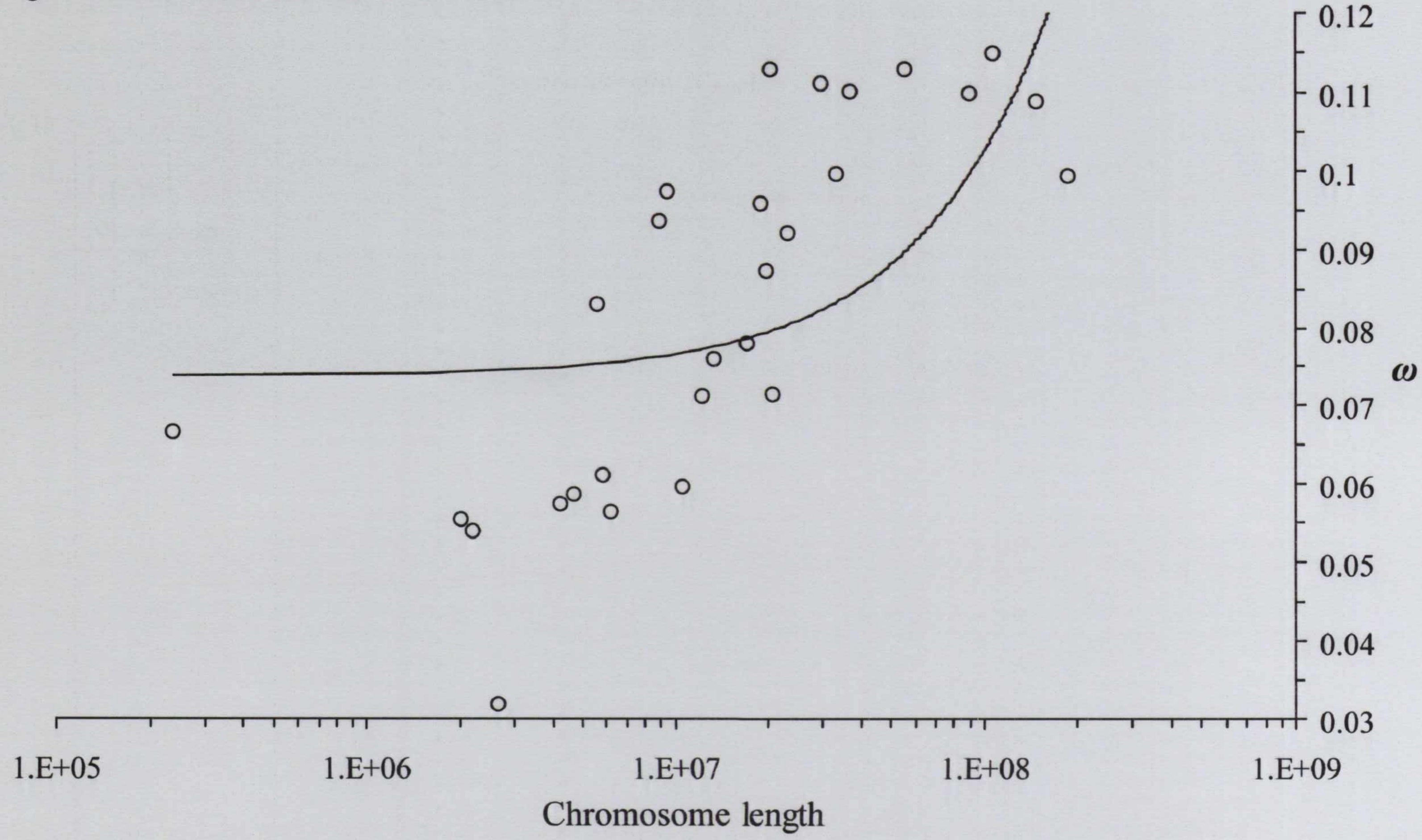
Increasing chicken chromosome size correlated with higher chromosomal GC3 rates for chicken genes ($r^2 = 0.435$, p = 0.010; Figure 2.2) and their zebra finch orthologs ($r^2 = 0.358$, p = 0.030). Although smaller chromosomes tended to have lower chromosomal ω values for all genes ($r^2 = 0.325$, p = 0.046; Figure 2.3), they had a higher frequency of genic SNPs per kb due to a higher incidence of genes (Figure 2.4). This was consistent with previous human-chicken comparison (International Chicken Genome Sequencing Consortium 2004) and analyses of SNP diversity (International Chicken Polymorphism Map Consortium 2004) and supported the relative non-biased nature of the ortholog datasets. Further F-tests involving chromosomal categories binned in groups according to size suggested that the manner in which these were previously assigned has produced artefactual results (International Chicken Genome Sequencing Consortium 2004); unbinned chromosomes allowed a more robust analysis.

Figure 2.2. Correlation of chicken chromosome length with chromosomal GC3 content for chicken genes and their zebra finch orthologs.



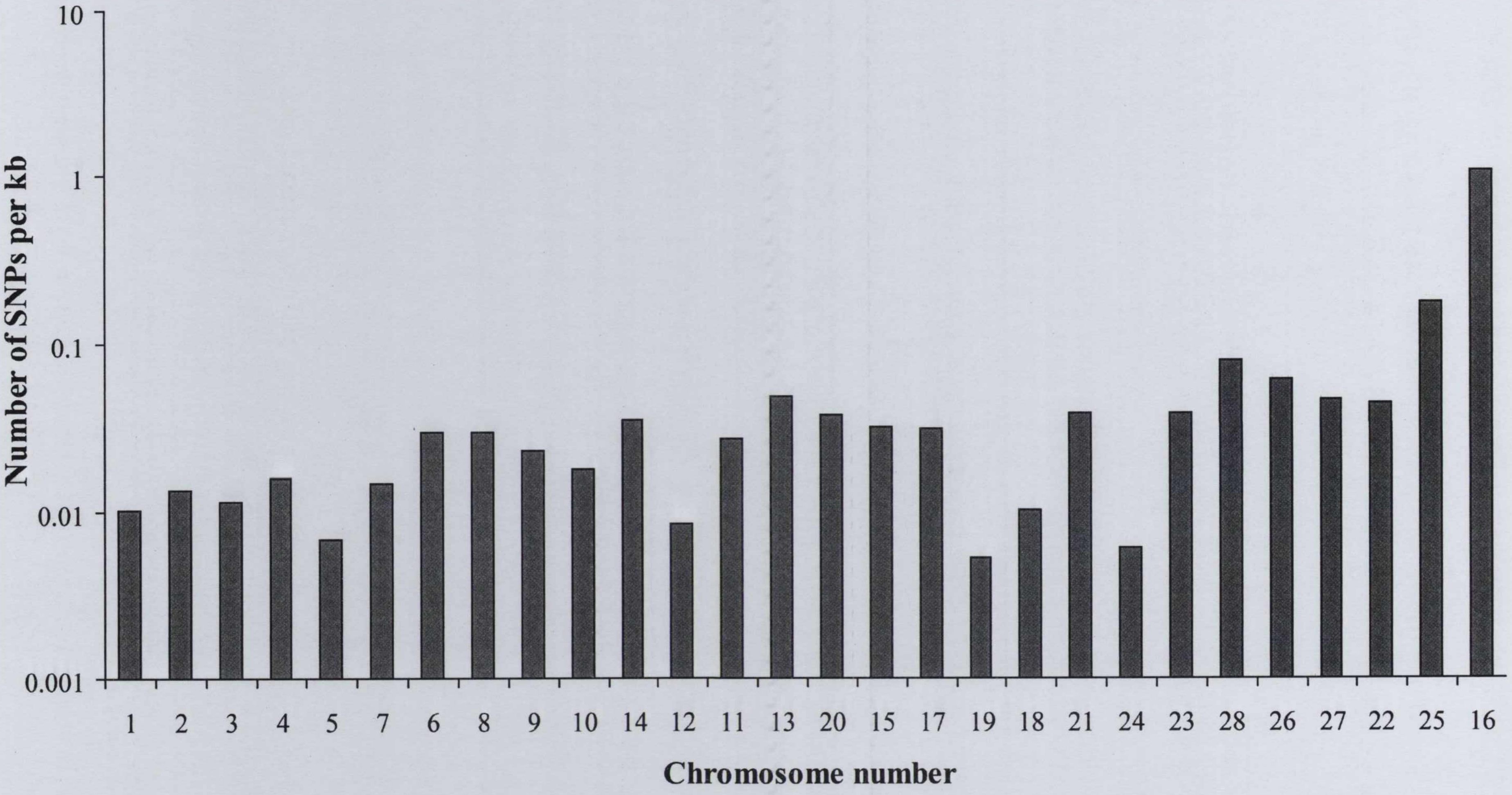
The best fitting linear correlations of GC content at the third codon position (GC3) for chicken (red, $r^2 = 0.435$, $p = 0.010$) and zebra finch (blue, $r^2 = 0.358$, $p = 0.030$) with chicken chromosome size (on a log scale) are shown by the lines.

Figure 2.3. Correlation of chicken chromosome size with ω.



The best fitting linear correlation of chromosome length with chromosomal rates of $\omega = d_N/d_S$ is shown by the line ($r^2 = 0.325$, $p = 0.046$).

Figure 2.4. Number of SNPs per kb of chicken transcript sequence covered for each chromosome ordered according to decreasing size.



3’ and 5’ UTR, indel, frameshift, upstream, downstream, splice site, intronic, exonic and stop-codon SNPs were included.

2.4 Discussion

This study combined an intraspecific analysis of chicken variation and an interspecies survey of chicken and zebra finch genes with orthologous human functions. It demonstrated that amino-acid changing sites in immune genes are subject to purifying selection both in chicken and the avian lineage. This demonstrated that there was no evidence of a significant relaxation of the selective constraint on chicken immune genes as a group since domestication.

In spite of this, chicken immune genes have undergone a higher ancestral rate of fixation of replacement substitutions than non-immune genes, symptomatic of a greater rate of directional selection (McDonald & Kreitman 1991). This was supported by the high proportion of amino acid changes fixed in chicken for immune genes. A previous study of chicken and zebra finch genes expressed in the brain estimated the portion of nonsynonymous polymorphisms in chicken that were fixed by positive selection (0.20; Axelsson & Ellegren 2009), indicating that immune genes as a group are under a greater frequency of selective events. The negative α value for non-immune genes indicated the incidence of deleterious variants on the chicken-zebra finch lineage (Eyre-Walker 2006), which is backed by evidence that a substantial minority (0.23) of amino acid changes segregating in chicken are deleterious (Axelsson & Ellegren 2009).

The considerable conservation of nonsynonymous sites at immune genes within chickens has probably exaggerated the perceived strength of positive selection on these sites on the avian lineage (Smith & Eyre-Walker 2002). Additionally, it is possible that high recombination or resequencing of rare polymorphisms may inflate this figure (Axelsson & Ellegren 2009), and while the chicken's high variability suggests that it has not gone through a major population bottleneck since domestication (International Chicken Polymorphism Map Consortium 2004), the fixation of deleterious alleles in tandem with population size increases can amplify estimates of α (Schlenke & Begun 2003). Nonetheless, the fraction of fixed replacement substitutions that were under positive selection at chicken immune genes further supports the assertion that this functional category was historically subject to stronger adaptive forces from pathogens and consequently undergoes directional selective sweeps more frequently than other gene groups (Schlenke & Begun 2003).

McDonald-Kreitman tests suggested that 26 genes were under pervasive purifying selection within chicken. As a group, they had significantly reduced GC content, which is associated with reduced variation (International Chicken Genome Sequencing Consortium 2004), and two of these genes were associated with recombination. Lower GC content is associated with decreased recombination (International Chicken Polymorphism Map Consortium 2004) implying that the impact of recombination on diversity may necessitate modification of genes controlling this process.

A further examination of GC content showed that it was substantially lower in zebra finch compared to chicken, and significantly higher in immune genes. Chromosome size appeared to be related to ω values, suggesting that genes on larger chromosomes may evolve faster, as has been suggested previously (Schlenke & Begun 2003, Axelsson et al. 2005). Once robust chromosomal assignments of zebra finch genes are established, this could be explored further in order to understand the complex patterns of chromosomal fission, fusion and rearrangements in avian species (Hannson et al. 2009, Nie et al. 2009, Griffin et al. 2007, Stapley et al. 2008, Itoh & Arnold 2005) and how this relates to GC content and the evolutionary dynamics of immune genes. Additionally, further sequencing and annotation of the zebra finch and other bird genomes will allow a more comprehensive testing of selection operating in the chicken genome.

Submitted paper

This chapter formed the basis for a submitted manuscript to BMC Research Notes in 2009 (in review since the 14th of May) entitled “Evidence of the adaptive evolution of immune genes in chicken”. The authors are: Downing T, Cormican P, O’Farrelly C, Bradley DG and Lloyd AT.

CHAPTER 3

Identifying genomic variation in chicken expressed sequence tags

3.1 Introduction

Examining diversity across the chicken genome can inform on patterns of variation at genes of interest: investigating variability in ESTs is one such approach to explore this. ESTs are short cDNA sequences derived from randomly selected clones in a DNA library (Boguski et al. 1993). Advantages of using ESTs include the high number in which they are produced and that this expression level is related to the tissue which was sampled. Clustered sets of ESTs can reveal variable sites in genes, novel genes and splicing variants; however, this is compromised by the sequence quality of ESTs, which is generally undetermined or poor (Li et al. 2009).

3.1.1 Nucleotide variation in the chicken genome:

Previous studies of chicken diversity have uncovered trends of high diversity at nucleotide sites (International Chicken Genome Sequencing Consortium 2004). 2.8 million SNPs were identified by comparing a 6.6x coverage reference red JF genome to 0.25x coverage of non-overlapping genomes of commercial broiler, layer and silkie chickens (International Chicken Polymorphism Map Consortium 2004). 1,210 SNPs were found in 23,427 chicken ESTs using an approach incorporating the visual screening of chromatogram traces, a method traditionally used for resequenced data (Kim et al. 2004). ESTs have been effective in analysis of diversity in other domestic organisms as well (for example, Hawken et al. 2004).

Consequently, studying genomic variability in chicken ESTs is likely to illuminate the diversity in the genome. SNPs in ESTs from genes associated with disease may have consequences for the immune response function and the chicken evolutionary history since domestication (International Chicken Genome Sequencing Consortium 2004).

In this study, a set of clustered GenBank ESTs was reciprocally searched against and aligned with chicken genes so that valid SNPs could be identified while excluding EST sequencing errors. Orthologous human functions and processes were assigned to the chicken genes and a web-based database of EST alignments was developed. SNPs discovered in ESTs were subsequently investigated in two resequenced candidate genes: lysozyme and toll-like receptor-1like A (TLR1LA).

3.2 Methods

3.2.1 Development of a robust EST SNP dataset

All available chicken EST sequences (578,354) were downloaded from GenBank dbEST (www.ncbi.nlm.nih.gov/dbEST/) and cleaned of vector contaminants using SeqClean. Repetitive sequences were masked out using RepeatMasker. Tgicl (Pertea et al. 2003) clustered the ESTs into 52,718 consensus contigs conservatively assuming a length ≥ 100 bases and an identity ≥ 96% for overlapping regions (Figure 3.1).

In order to identify orthologous regions between the EST contigs and chicken coding sequences (CDS), chicken Refseq mRNA sequences (18,039) downloaded from GenBank were aligned to the EST contigs in a reciprocal manner using Blastn (Altschul et al. 1990). Subject to conditions of having an E value $\leq 10^{-3}$, a length > 100 bp and an identity > 98% to remove incorrect sequences, this approach identified 15,755 best hits of mRNAs on ESTs, 31,870 best hits of ESTs on mRNAs, which ultimately yielded 12,396 reciprocal best hit (RBH) pairs.

A series of Perl scripts were implemented to correct the orientation of the RBHs (run_analysis3.pl) and to obtain their sequence properties: the sequence length, number of polymorphisms, sequence positions, and EST coverage for each variable site. The program Cap3 (Huang & Maddan 1999) was used to trim out low-quality regions and to perform alignments to identify RBHs where there were at least four ESTs clustered with a CDS sequence.

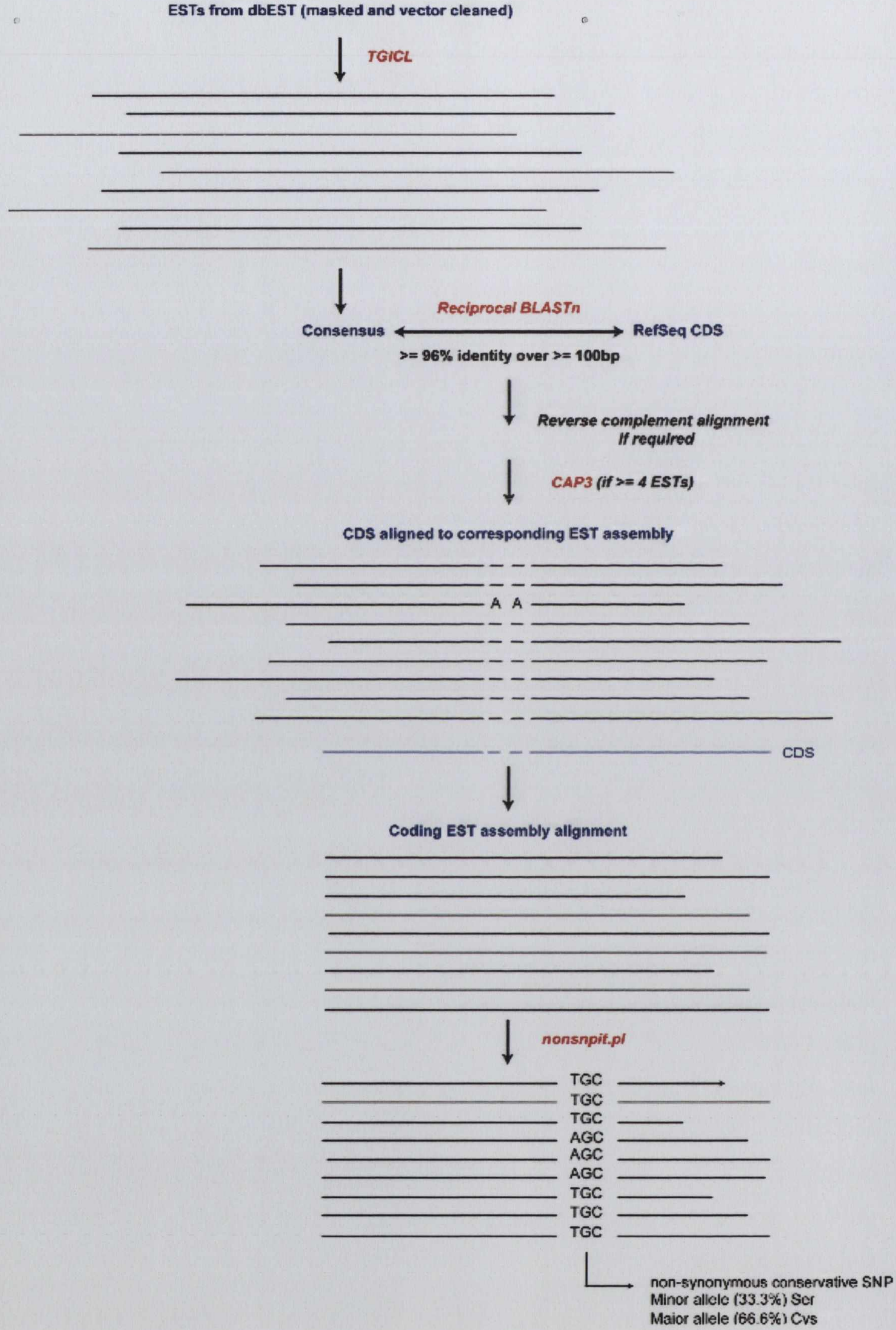
3.2.2 Identifying valid and functional EST SNPs

Valid RBH pairs were aligned and trimmed using Perl including a script (chick_nonSNPit.pl) that parsed the sequences into the correct protein-coding codon frame so that nonsynonymous and synonymous SNPs could be identified. Using chicken RefSeq genes as the reference sequences, it was possible to determine the frequency of the derived and ancestral alleles in the ESTs. The proportion of nonsynonymous (P_N) to synonymous (P_S) amino acid changes were determined. Given a prior expectation of a P_N/P_S ratio of 1 (Yang 2002), silent changes are expected to occur at a neutral rate deviations from this where $P_N > P_S$ might indicate a relative surfeit of nonsynonymous substitutions in EST sequences. Alternatively, if $P_S > P_N$ this suggests more conservation at nonsynonymous sites. In addition, the relative incidences of non-conservative (non-con) to conservative (con) amino acid changes were ascertained (non-con/con) using protein impact prediction software (SIFT; http://blocks.fhcrc.org/sift/; Ng & Henikoff 2003): a high frequency of non-conservative substitutions could be a sign of functional relevance (International Chicken Genome Sequencing Consortium 2004). SIFT works by aligning the sequence of interest using PSI-BLAST (Altschul et al. 1997) with similar sequences, which were compiled from a Uniref database (Apweiler et al. 2004) for 2 iterations to determine the best hits: from these SIFT infers the substitution impact scores.

The 18,039 chicken Refseq mRNAs were aligned with human GenBank protein refseqs (34,065) using Blastp (Altschul et al. 1990) to assign orthologous human functional categories from Panther (www.pantherdb.org, Thomas et al. 2003) where the E value $\geq 10^{-10}$ for the chicken genes. From these, a set of EST SNPs in immune genes was established. Mutations in ESTs causing stop codons were examined in particular because of their abrupt consequences for gene function and implications for deleterious disease. Chromatograms of ESTs on GenBank for genes with EST SNPs causing stop codons that had adequate base coverage were examined for validity and other GenBank databases (Homologene, PubMed, Entrez) were checked for information pertaining to variability at this gene.

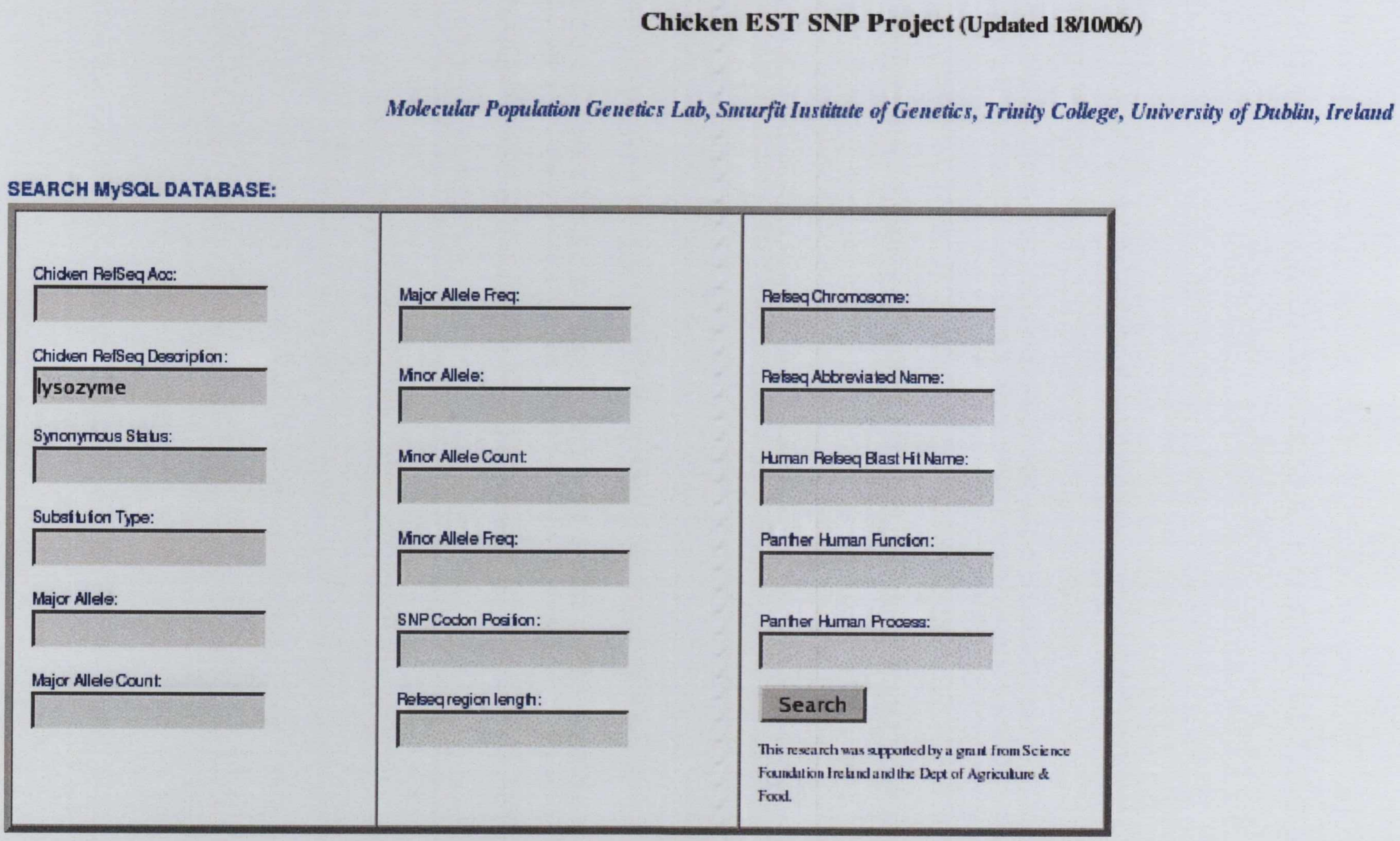
Additional Perl scripts were used to assemble all aligned RBH gene ESTs (run_analysis_nocap3.pl). A local interactive web server was developed in HTML so that the properties of each pair could be searched locally using MySQL software, printed onscreen and the alignment could be graphically visualised with Jalview, a Java alignment editor (Clamp et al 2004). Searchable terms included GenBank accession number, gene names, SNP type and position, major and minor allele type, count and frequency, gene CDS length, chromosome, human ortholog, human ortholog function and process (Figure 3.2). These were implemented with pnpsdets.pl and genbankdets.pl (see Appendix A). Jalview displayed the local degree of sequence conservation and quality, which helped determine the validity of SNPs.

Figure 3.1. Analysis pipeline for identifying SNPs in chicken ESTs.


ESTs from dbEST (masked and vector cleaned)
TGICL
Reciprocal BLASTn
Consensus
RefSeq CDS
>= 96% identity over >= 100bp
Reverse complement alignment if required
CAP3 (if >= 4 ESTs)
CDS aligned to corresponding EST assembly
A A
CDS
Coding EST assembly alignment
nonsnpit.pl
TGC
TGC
TGC
AGC
AGC
AGC
TGC
TGC
TGC
non-synonymous conservative SNP
Minor allele (33.3%) Ser
Major allele (66.6%) Cys

Programs are in red, steps are in blue and parameters are in black.

Figure 3.2. Chicken EST SNP database search entry page.



Chicken EST SNP Project (Updated 18/10/06/)

Molecular Population Genetics Lab, Smurfit Institute of Genetics, Trinity College, University of Dublin, Ireland

SEARCH MySQL DATABASE:

Chicken RefSeq Acc:

Chicken RefSeq Description: lysozyme

Synonymous Status:

Substitution Type:

Major Allele:

Major Allele Count:

Major Allele Freq:

Minor Allele:

Minor Allele Count:

Minor Allele Freq:

SNP Codon Position:

Refseq region length:

Refseq Chromosome:

Refseq Abbreviated Name:

Human Refseq Blast Hit Name:

Panther Human Function:

Panther Human Process:

This research was supported by a grant from Science Foundation Ireland and the Dept of Agriculture & Food.

3.3 Results

In order to explore diversity in the chicken genome, EST sequences were reciprocally aligned with a reference chicken gene dataset. This approach identified 3,154 genes whose coding regions had at least one novel SNP present in four or more ESTs.

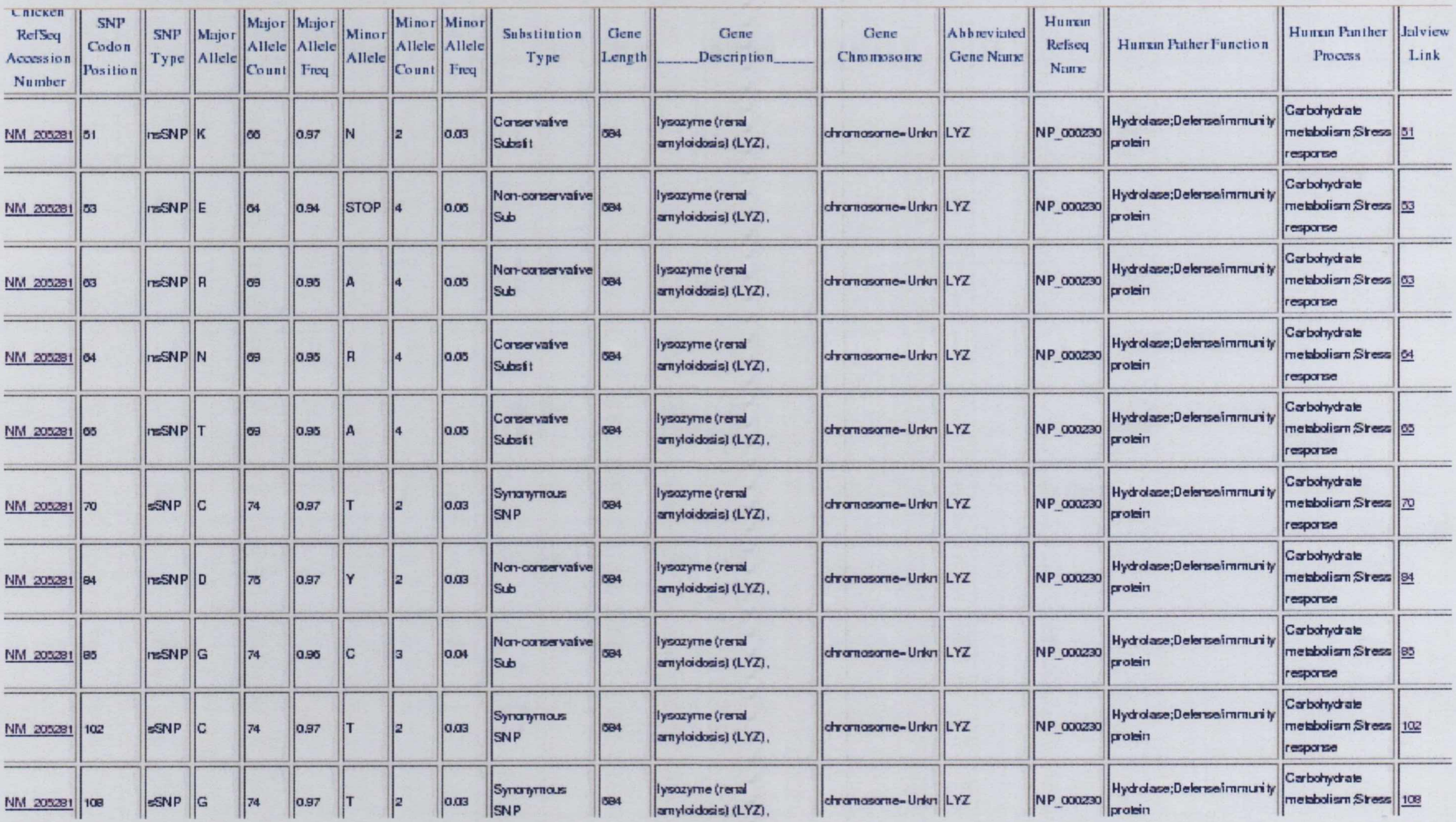
3.3.1 EST SNP database:

A web server was created that listed the gene names, accession number, length, SNP amino acid position and type, major and minor allele type, frequency and number, chromosome, human ortholog name, function and process (Figure 3.3). In addition, a link to Jalview was provided so that the SNPs and alignments could be visually examined. This software generated a Java applet with the GenBank and EST accession numbers, sequence and SNP positions, local sequence conservation, quality and consensus in a graphical format that enabled perusal of genes and SNPs of interest (Figure 3.4).

3.3.2 EST SNP discovery:

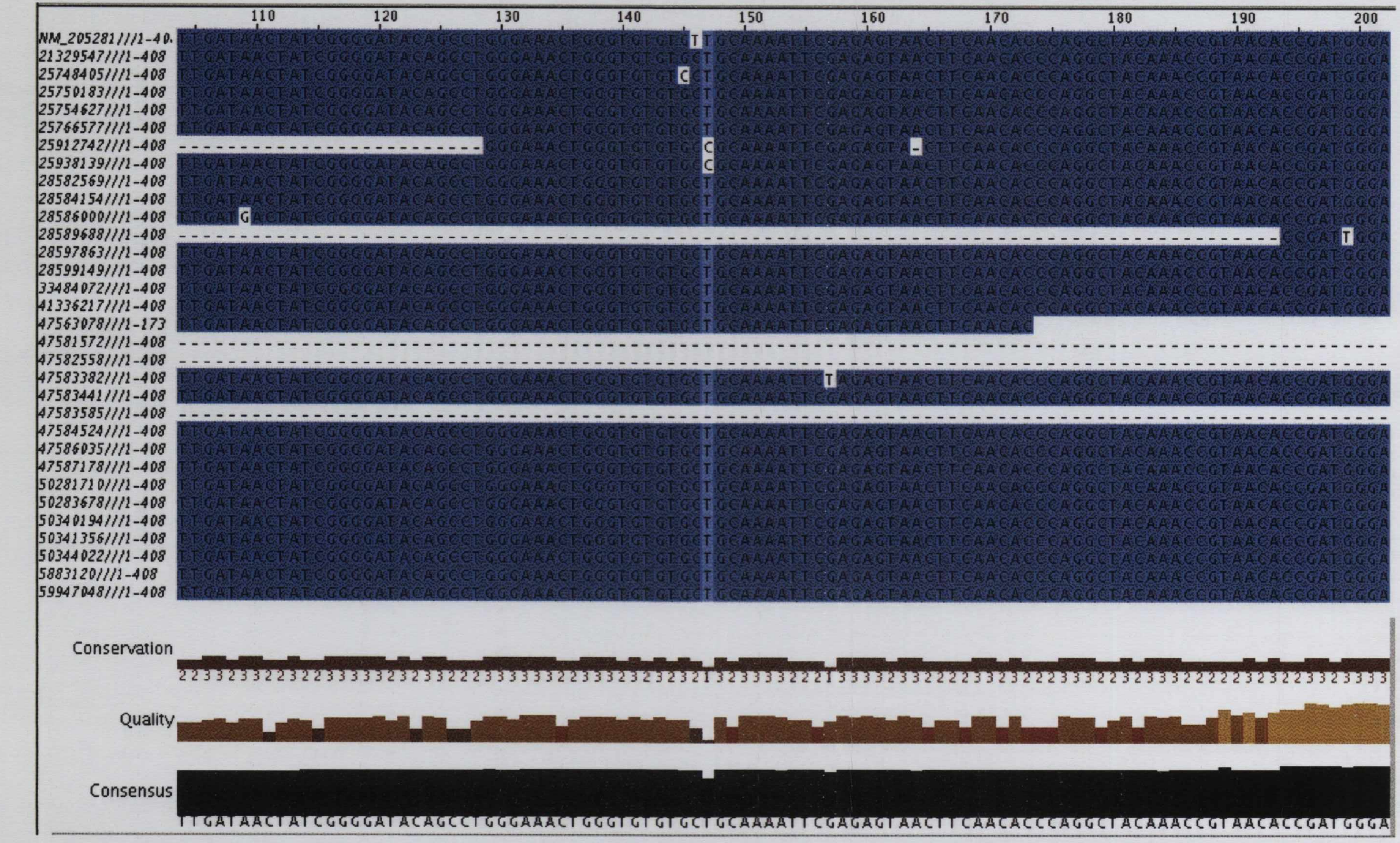
The reciprocal alignment of chicken genes with ESTs found 3,154 best hit pairs with at least one EST SNP. This identified 8,630 SNPs, 5,673 of which were synonymous. 2,885 nonsynonymous SNPs were discovered, 1,896 of which were conservative substitutions and 989 of which were non-conservative – including 79 SNPs encoding stop codons. Genes with high P_N relative to P_S and non-conservative (non-con) compared to conservative (con) changes were identified: 1,296 genes had $P_N/P_S \geq 0.5$ or non-con/con ≥ 0.5 (see Appendix B for details). 560 genes had $P_N/P_S \geq 1$ and 264 had more non-conservative compared to conservative changes (non-con/con > 1). Those with the highest P_N values were collated: 20 genes had $P_N > 10$ (Table 3.1): these included a pancreatic amylase (AMY2A), a haemoglobin gene (HBA1) and lactate dehydrogenase B (LDHB). Most of these genes had a high number of observed EST SNPs, $P_N/P_S > 1$ and non-con/con ≥ 0.5.

Figure 3.3. Screenshot of web server results for such for lysozyme gene EST



Chicken RefSeq Accession Number	SNP Codon Position	SNP Type	Major Allele	Major Allele Count	Major Allele Freq	Minor Allele	Minor Allele Count	Minor Allele Freq	Substitution Type	Gene Length	Gene Description	Gene Chromosome	Abbreviated Gene Name	Human Refseq Name	Human Pather Function	Human Panther Process	Jalview Link
NM_205281	51	nsSNP	K	66	0.97	N	2	0.03	Conservative Subsit	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress response	51
NM_205281	53	nsSNP	E	64	0.94	STOP	4	0.06	Non-conservative Sub	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress response	53
NM_205281	63	nsSNP	R	69	0.95	A	4	0.05	Non-conservative Sub	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress response	63
NM_205281	64	nsSNP	N	69	0.95	R	4	0.05	Conservative Subsit	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress response	64
NM_205281	65	nsSNP	T	69	0.95	A	4	0.05	Conservative Subsit	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress response	65
NM_205281	70	sSNP	C	74	0.97	T	2	0.03	Synonymous SNP	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress response	70
NM_205281	84	nsSNP	D	75	0.97	Y	2	0.03	Non-conservative Sub	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress response	84
NM_205281	85	nsSNP	G	74	0.96	C	3	0.04	Non-conservative Sub	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress response	85
NM_205281	102	sSNP	C	74	0.97	T	2	0.03	Synonymous SNP	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress response	102
NM_205281	108	sSNP	G	74	0.97	T	2	0.03	Synonymous SNP	684	lysozyme (renal amyloidosis) (LYZ),	chromosome=Unkn	LYZ	NP_000230	Hydrolase;Defenseimmunity protein	Carbohydrate metabolism;Stress	108

Figure 3.4. Screenshot of Jalview alignment window for the lysozyme gene.



Legend to Figure 3.4: The uppermost sequence shown is the genome reference sequence. Dashes represent EST regions with no determined sequence. White bases represent SNPs. Blue bases represent conserved sequence. Sequence conservation, quality and consensus increase correspondingly with bar height. Codon E53 is represented by bases 157, 158 and 159 – only one of the four ESTs with stop SNPs is visually displayed in this format due to the presence of a large number of ESTs (70), most of which are not shown due to page size constraints.

Table 3.1. Genes with the highest P_N values.

GenBank Gene Name	Accession Number	P_N	P_S	P_N/P_S[1]	non-con	con	non-con/con	SNPs[2]
haemoglobin alpha 1, HBA1	NM_001004376	26	5	5.200	8	18	0.444	32
amylase, alpha 2A; pancreatic, AMY2A	NM_001001473	23	5	4.600	9	14	0.643	29
ribosomal protein S6, RPS6	NM_205225	23	4	5.750	10	13	0.769	28
60S ribosomal protein L8	XM_416772	23	10	2.300	9	14	0.643	34
ferritin heavy polypeptide 1, FTH1	NM_205086	22	5	4.400	13	9	1.444	28
elastase 2A, ELA2A	NM_001032390	21	13	1.615	9	12	0.750	35
apolipoprotein A-I, APOA1	NM_205525	19	11	1.727	7	12	0.583	31
protease serine 2, trypsin 2, PRSS2	NM_205384	18	11	1.636	10	8	1.250	30
eukaryotic translation elongation factor 1 alpha 1	NM_204157	18	5	3.600	8	10	0.800	24
ribosomal protein L7a, RPL7A	NM_001004379	17	6	2.833	5	12	0.417	24
heat shock cognate 70, HSC70	NM_205003	17	5	3.400	6	11	0.545	23
lactate dehydrogenase B, LDHB	NM_204177	16	7	2.286	3	13	0.231	24
polyubiquitin, LOC417602	XM_415847	15	11	1.364	7	8	0.875	27
40S ribosomal protein S10	XM_418029	15	0	high	8	7	1.143	16
glyceraldehyde-3-phosphate dehydrogenase, GAPDH	NM_204305	14	6	2.333	8	6	1.333	21
ribosomal protein L7, RPL7	NM_001006345	12	4	3.000	3	9	0.333	17
DMRT1 isoform e, LOC395181	XM_418817	12	4	3.000	5	7	0.714	17
ribosomal protein P0-like	XM_425751	11	4	2.750	5	6	0.833	16
40S ribosomal protein S2	XM_414845	11	5	2.200	3	8	0.375	17
DKFZp434C0328 protein	XM_416574	11	10	1.100	5	6	0.833	22

[1] P_N/P_S is defined as "high" where $P_S = 0$ and $P_N > 0$. [2] Total number of SNPs in ESTs observed at the locus.

Genes with more than four non-conservative substitutions were assembled (Table 3.2) – again this included AMY2A and HBA1. As before, this set of genes was highly variable and displayed high P_N/P_S values. The group of 20 genes with the highest P_N/P_S values where $P_S > 0$ (Table 3.3) included the CD3ε gene, which is immunity-related (Göbel & Fluri 1997). With 23 and 26 nonsynonymous substitutions respectively, AMY2A and HBA1 formed part of this dataset. Examining genes with high ratios of non-conservative to conservative substitutions where there was at least one conservative SNP did not uncover any known genes, perhaps because the low incidence of conservative amino acid changes excluded many genes (Table 3.4).

Table 3.2. Genes with the highest number of non-conservative substitutions.

GenBank Gene Name	Accession Number	P_N	P_S	P_N/P_S [1]	non-con	con	non-con/con	SNPs [2]
ferritin heavy polypeptide 1, FTH1	NM_205086	22	5	4.400	13	9	1.444	28
protease serine 2, trypsin 2, PRSS2	NM_205384	18	11	1.636	10	8	1.250	30
ribosomal protein S6, RPS6	NM_205225	23	4	5.750	10	13	0.769	28
amylase, alpha 2A; pancreatic, AMY2A	NM_001001473	23	5	4.600	9	14	0.643	29
elastase 2A, ELA2A	NM_001032390	21	13	1.615	9	12	0.750	35
60S ribosomal protein L8	XM_416772	23	10	2.300	9	14	0.643	34
glyceraldehyde-3-phosphate dehydrogenase, GAPDH	NM_204305	14	6	2.333	8	6	1.333	21
eukaryotic translation elongation factor 1 alpha 1	NM_204157	18	5	3.600	8	10	0.800	24
40S ribosomal protein S10	XM_418029	15	0	high	8	7	1.143	16
hemoglobin alpha 1, HBA1	NM_001004376	26	5	5.200	8	18	0.444	32
polyubiquitin, LOC417602	XM_415847	15	11	1.364	7	8	0.875	27
apolipoprotein A-I, APOA1	NM_205525	19	11	1.727	7	12	0.583	31
heat shock cognate 70, HSC70	NM_205003	17	5	3.400	6	11	0.545	23
NECAP endocytosis associated 2, NECAP2	NM_001012837	8	2	4.000	5	3	1.667	11
ribosomal protein P0-like	XM_425751	11	4	2.750	5	6	0.833	16
DMRT1 isoform e, LOC395181	XM_418817	12	4	3.000	5	7	0.714	17
ribosomal protein L7a, RPL7A	NM_001004379	17	6	2.833	5	12	0.417	24
DKFZp434C0328 protein, LOC41	XM_416574	11	10	1.100	5	6	0.833	22
hypothetical protein FLJ2356	XM_416107	6	1	6.000	5	1	5.000	8
ribosomal protein L4, RPL4	NM_001007479	8	7	1.143	5	3	1.667	16

[1] P_N/P_S is defined as “high” where $P_S = 0$ and $P_N > 0$. [2] Total number of SNPs in ESTs observed at the locus.

It should be noted that protein impact predictions by SIFT were not sufficiently confident due to excessive protein sequence divergence between chicken and mammals for most RBH pairs, thus substitutions could not be assigned as neutral or deleterious, though the substitution outcome (conservative and non-conservative) could be inferred. The genes with the most EST SNPs included those with the most nonsynonymous changes, such as AMY2A, HBA1 and LDHB (Table 3.5).

Table 3.3. Genes with the highest P_N/P_S values where $P_S > 0$.

GenBank Gene Name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs [1]
CD3E antigen epsilon polypeptide, TiT3 complex,	NM_206904	8	1	8.000	3	5	0.600	10
hypothetical protein MGC3508	XM_420702	7	1	7.000	0	7	0.000	9
hypothetical gene supported by CR353961	XM_429985	6	1	6.000	4	2	2.000	8
hypothetical protein FLJ2356	XM_416107	6	1	6.000	5	1	5.000	8
tyrosine 3-monooxygenase/tryptophan 5-monooxygenas	NM_001006219	6	1	6.000	3	3	1.000	8
DAZ associated protein 1, LOC427266	NM_001031428	6	1	6.000	2	4	0.500	8
ribosomal protein S6, RPS6	NM_205225	23	4	5.750	10	13	0.769	28
hemoglobin alpha 1, HBA1	NM_001004376	26	5	5.200	8	18	0.444	32
soc-2 suppressor of clear homolog, C. elegans, SH	NM_001031236	5	1	5.000	0	5	0.000	7
Secernin 2, LOC425759 part	XM_423480	5	1	5.000	4	1	4.000	7
zinc finger FYVE domain	XM_421128	5	1	5.000	1	4	0.250	7
chromosome 1 open reading frame	XM_422229	5	1	5.000	3	2	1.500	7
AQ, LOC395744	NM_204914	5	1	5.000	3	2	1.500	7
amylase alpha 2A; salivary, AMY2A	NM_001001473	23	5	4.600	9	14	0.643	29
ferritin heavy polypeptide 1, FTH1	NM_205086	22	5	4.400	13	9	1.444	28
NECAP endocytosis associated 2, NECAP2	NM_001012837	8	2	4.000	5	3	1.667	11
Catalase, LOC423601 partial	XM_421487	4	1	4.000	2	2	1.000	6
hypothetical protein BC011840, LOC42608	NM_001031355	4	1	4.000	2	2	1.000	6
hypothetical protein FLJ2062	XM_419675	4	1	4.000	1	3	0.333	6
NADH-cytochrome b5 reductase, CYB5R	XM_420957	4	1	4.000	0	4	0.000	6

[1] Total number of SNPs in ESTs observed at the locus.

Table 3.4. Genes with the highest ratio of non-conservative to conservative substitutions where there is at least one conservative change.

GenBank Gene Name	Accession Number	P_N	P_S	P_N/P_S [1]	non-con	con	non-con/con	SNPs [2]
hypothetical protein FLJ2356	XM_416107	6	1	6.000	5	1	5.000	8
ribosomal protein L29	XM_425143	5	0	high	4	1	4.000	6
Secernin 2, LOC425759 part	XM_423480	5	1	5.000	4	1	4.000	7
cystatin C, CST3	NM_205500	5	0	high	4	1	4.000	6
LOC426122	XM_430363	4	2	2.000	3	1	3.000	7
hypothetical gene supported by BX933436	XM_429924	4	2	2.000	3	1	3.000	7
CUG triplet repeat binding protein 2, CUGBP2	NM_204260	4	1	4.000	3	1	3.000	6
ubiquinol--cytochrome c reductase	XM_414356	4	2	2.000	3	1	3.000	7
Ras association, RalGDS/AF-6 domain family 2, RAS	NM_001030884	4	2	2.000	3	1	3.000	7
TIMP metallopeptidase inhibitor 3, Sorsby fundus	NM_205487	7	6	1.167	5	2	2.500	14
ENSANGP00000017034, LOC42322	XM_421148	3	1	3.000	2	1	2.000	5
mitochondrial ribosomal protein	XM_420108	3	3	1.000	2	1	2.000	7
B6.1, LOC396098	NM_205182	6	3	2.000	4	2	2.000	10
hypothetical gene supported by BX931271	XM_417971	3	0	high	2	1	2.000	4
RIKEN cDNA 2400003L07	XM_422733	3	2	1.500	2	1	2.000	6
hypothetical protein FLJ22626, LOC42236	NM_001006437	3	0	high	2	1	2.000	4
hypothetical gene supported by CR353961	XM_429985	6	1	6.000	4	2	2.000	8
hypothetical gene supported by BX933825	XM_429587	3	2	1.500	2	1	2.000	6
mitochondrial ATP synthase	XM_416717	3	2	1.500	2	1	2.000	6
hematological and neurological expressed 1, HN1	NM_001006425	3	0	high	2	1	2.000	4

[1] P_N/P_S is defined as "high" where $P_S = 0$ and $P_N > 0$. [2] Total number of SNPs in ESTs observed at the locus.

Table 3.5. Genes with the highest number of EST SNPs.

GenBank Gene Name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
elastase 2A, ELA2A	NM_001032390	21	13	1.615	9	12	0.750	35
60S ribosomal protein L8	XM_416772	23	10	2.300	9	14	0.643	34
hemoglobin alpha 1, HBA1	NM_001004376	26	5	5.200	8	18	0.444	32
apolipoprotein A-I, APOA1	NM_205525	19	11	1.727	7	12	0.583	31
protease serine 2, trypsin 2, PRSS2	NM_205384	18	11	1.636	10	8	1.250	30
amylase alpha 2A; salivary, AMY2A	NM_001001473	23	5	4.600	9	14	0.643	29
ferritin heavy polypeptide 1, FTH1	NM_205086	22	5	4.400	13	9	1.444	28
ribosomal protein S6, RPS6	NM_205225	23	4	5.750	10	13	0.769	28
polyubiquitin, LOC417602	XM_415847	15	11	1.364	7	8	0.875	27
ribosomal protein L7a, RPL7A	NM_001004379	17	6	2.833	5	12	0.417	24
eukaryotic translation elongation factor 1 alpha 1	NM_204157	18	5	3.600	8	10	0.800	24
lactate dehydrogenase B, LDHB	NM_204177	16	7	2.286	3	13	0.231	24
heat shock cognate 70, HSC70	NM_205003	17	5	3.400	6	11	0.545	23
enolase 1, alpha, ENO1	NM_205120	9	12	0.750	2	7	0.286	22
DKFZp434C0328 protein, LOC41	XM_416574	11	10	1.100	5	6	0.833	22
glyceraldehyde-3-phosphate dehydrogenase, GAPDH	NM_204305	14	6	2.333	8	6	1.333	21
ribosomal protein S4, LOC396001	NM_205108	9	10	0.900	4	5	0.800	20
ribosomal protein S3, RPS3	NM_001030836	9	8	1.125	2	7	0.286	18
ribosomal protein L7, RPL7	NM_001006345	12	4	3.000	3	9	0.333	17
DMRT1 isoform e, LOC395181	XM_418817	12	4	3.000	5	7	0.714	17

[1] Total number of SNPs in ESTs observed at the locus.

3.3.3 Immunity-related gene EST SNPs:

Functions and processes of human Panther orthologs were assigned to chicken genes so that those associated with immunity could be identified. These numbered 79 genes with a total of 206 EST SNPs. 19 immune genes had $P_N/P_S \geq 0.5$ or non-con/con ≥ 0.5 and 4 of these had at least 3 nonsynonymous SNPs (nsSNPs): IRAK2, IL16, TLR1LA and TIMD4 (Table 3.6).

Although these 19 immune genes had more nonsynonymous (39) than synonymous (30) SNPs, only one third of these were non-conservative (13). Of TLR1LA's (GenBank accession number NM_001007488) 13 SNPs, four were nonsynonymous and two of these were non-conservative (Figure 3.5). At IL16, five of 11 SNPs in total were nonsynonymous and two of these were non-conservative. 10 SNPs were present at TIMD4, five of which were nonsynonymous and two of these were non-conservative. IRAK2 had seven SNPs – all of which were nonsynonymous but only 1 of these was a non-conservative change.

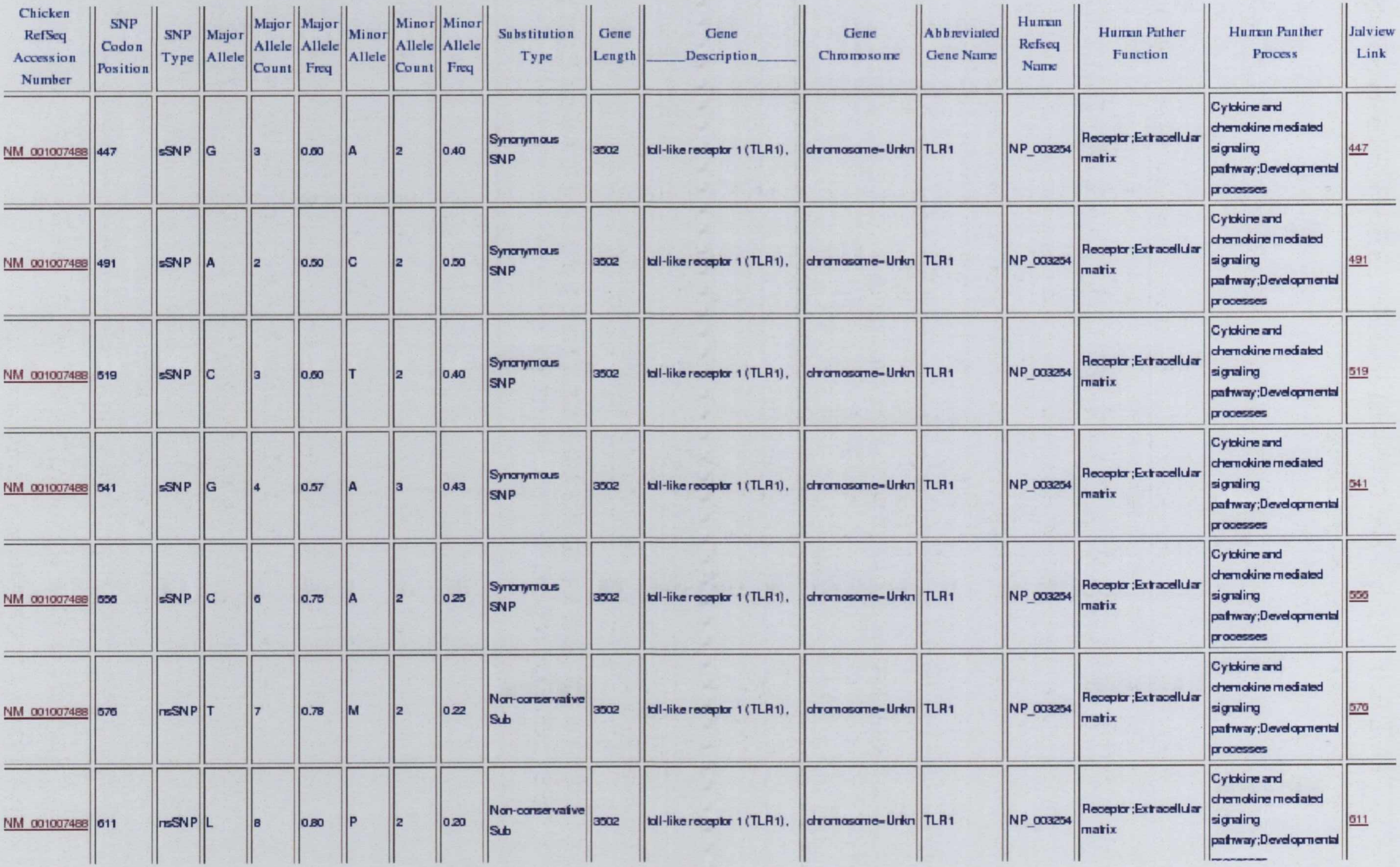
Table 3.6. EST SNPs at immunity-related genes with $P_N/P_S > 0.5$ or non-con/con ≥ 0.5.

Gene Name	GenBank[1]	P_N	P_S	P_N/P_S[2]	non-con	con	SNPs[3]
IL2Ra	NM_204596	1	1	1.000	0	1	2
B-defensin 10	NM_001001609	1	0	high	1	0	1
IL25	NM_001006342	1	0	high	0	1	1
IL16	NM_204352	5	6	0.833	2	3	11
Ig μ heavy chain	XM_428803	2	0	high	1	1	2
B-MA2	XM_415339	1	1	1.000	0	1	2
IRAK2	NM_001030605	7	0	high	1	6	7
IGHMBP2	NM_001031175	2	3	0.667	0	2	5
IRG1	NM_001030821	1	1	1.000	1	0	2
TNFRSF1B	NM_204439	1	0	high	0	1	1
TNFAIP8L1	NM_001006343	1	0	high	0	1	1
IL7R	XM_423732	1	0	high	1	0	1
N4BP2L2	NM_001012828	1	1	1.000	1	0	2
TIMD4	NM_001006149	5	5	1.000	2	3	10
TGIF1	NM_205379	1	1	1.000	0	1	2
IFRD1	NM_001001468	2	1	2.000	1	1	3
TLR1LA	NM_001007488	4	9	0.444	2	2	13
SEMA3F	NM_204258	1	1	1.000	0	1	3
NKRF	NM_001012887	1	0	high	0	1	2

[1] GenBank accession number. [2] P_N/P_S is defined as "high" where $P_S = 0$ and $P_N > 0$. [3] Total number of SNPs in ESTs observed at the locus.

TLR1LA was resequenced in a group commercial broilers, heritage chickens, and JF: red, grey, Ceylon and green. See Chapter 7 for further details on amplification, SNP detection and sequence analysis. Only one of the 13 EST SNPs at TLR1LA was detected during resequencing – this was at codon 491 in the single exon present at the gene (equivalent to base 1636 in the GenBank mRNA entry for TLR1LA and base 1872 in Chapter 7). This synonymous polymorphism was a C to A change encoding arginine. The ancestral C allele was present in the heritage birds, broilers, and red, grey, Ceylon and green JF. The derived A allele was observed in half of the 16 broiler genotypes. Both C and A nucleotides were present in 2 ESTs each: GenBank accession numbers AJ723681 and AJ723686 for C (Caldwell et al. 2005), and BU471924 and BU383051 for A (Boardman et al. 2002).

Figure 3.5. Screenshot of 1st 3 web server results for such for TLR1LA gene EST SNPs.



Chicken RefSeq Accession Number	SNP Codon Position	SNP Type	Major Allele	Major Allele Count	Major Allele Freq	Minor Allele	Minor Allele Count	Minor Allele Freq	Substitution Type	Gene Length	Gene Description	Gene Chromosome	Abbreviated Gene Name	Human Refseq Name	Human Pather Function	Human Panther Process	Jalview Link
NM_001007488	447	sSNP	G	3	0.60	A	2	0.40	Synonymous SNP	3502	toll-like receptor 1 (TLR1),	chromosome=Unkn	TLR1	NP_003254	Receptor;Extracellular matrix	Cytokine and chemokine mediated signaling pathway;Developmental processes	447
NM_001007488	491	sSNP	A	2	0.50	C	2	0.50	Synonymous SNP	3502	toll-like receptor 1 (TLR1),	chromosome=Unkn	TLR1	NP_003254	Receptor;Extracellular matrix	Cytokine and chemokine mediated signaling pathway;Developmental processes	491
NM_001007488	519	sSNP	C	3	0.60	T	2	0.40	Synonymous SNP	3502	toll-like receptor 1 (TLR1),	chromosome=Unkn	TLR1	NP_003254	Receptor;Extracellular matrix	Cytokine and chemokine mediated signaling pathway;Developmental processes	519
NM_001007488	541	sSNP	G	4	0.57	A	3	0.43	Synonymous SNP	3502	toll-like receptor 1 (TLR1),	chromosome=Unkn	TLR1	NP_003254	Receptor;Extracellular matrix	Cytokine and chemokine mediated signaling pathway;Developmental processes	541
NM_001007488	556	sSNP	C	6	0.75	A	2	0.25	Synonymous SNP	3502	toll-like receptor 1 (TLR1),	chromosome=Unkn	TLR1	NP_003254	Receptor;Extracellular matrix	Cytokine and chemokine mediated signaling pathway;Developmental processes	556
NM_001007488	576	nsSNP	T	7	0.78	M	2	0.22	Non-conservative Sub	3502	toll-like receptor 1 (TLR1),	chromosome=Unkn	TLR1	NP_003254	Receptor;Extracellular matrix	Cytokine and chemokine mediated signaling pathway;Developmental processes	576
NM_001007488	611	nsSNP	L	8	0.80	P	2	0.20	Non-conservative Sub	3502	toll-like receptor 1 (TLR1),	chromosome=Unkn	TLR1	NP_003254	Receptor;Extracellular matrix	Cytokine and chemokine mediated signaling pathway;Developmental	611

The current GenBank reference sequence (NM_001007488.3 from AADN02015855.1, a whole genome shotgun sequence) for TLR1LA has the ancestral C allele at this base, however, a previous version (NM_001007488.2 derived from AJ720806 from Caldwell et al. 2005) that was used in this analysis has the derived A allele (Figure 3.6).

3.3.4 EST SNPs coding for stop codons:

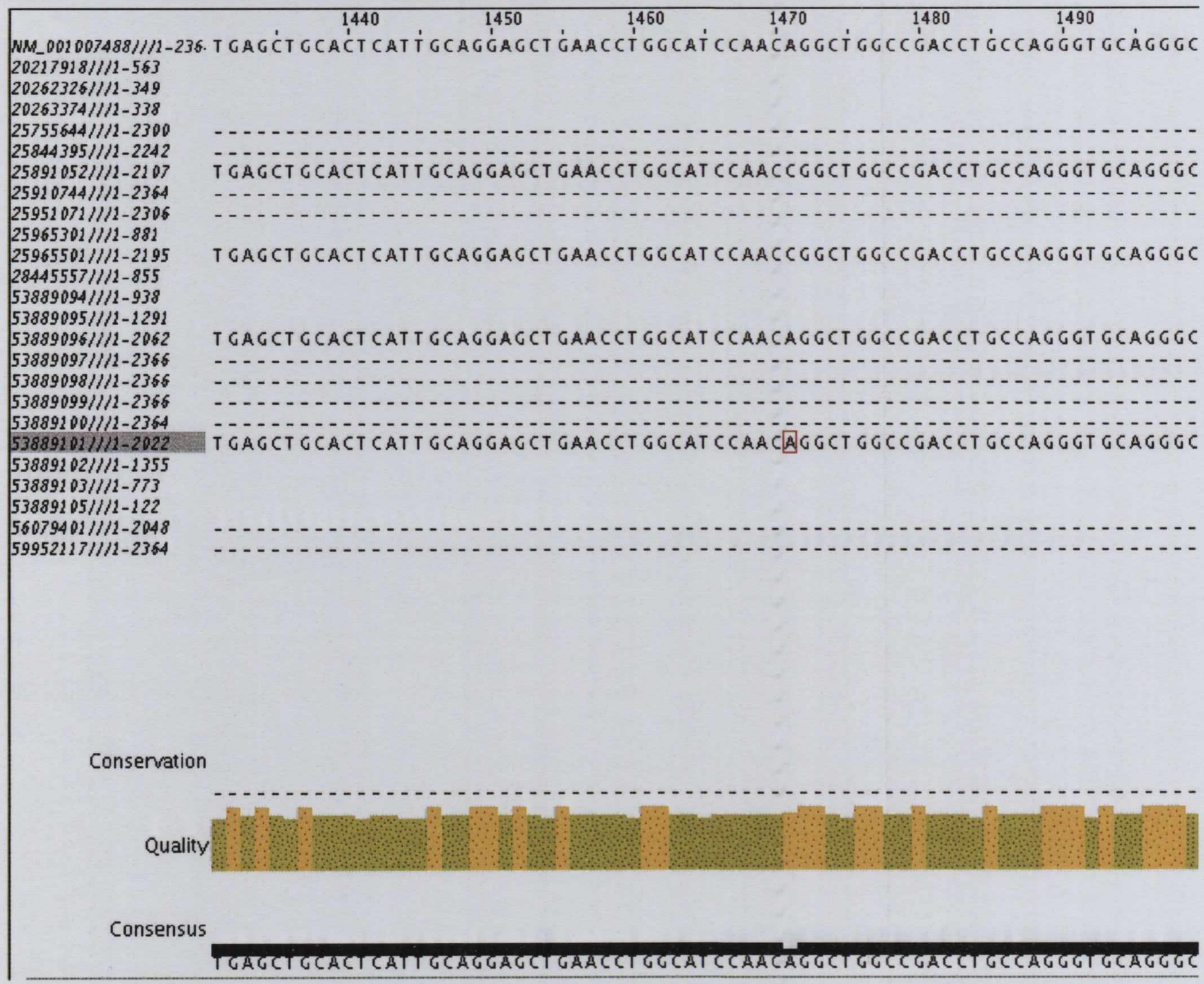
SNPs causing stop codons were examined in detail because of their functional implications for disease. Only one immune gene with such a SNP satisfying the criteria as set out above was identified: lysozyme. This innate immune gene had a stop SNP in 4 ESTs at the catalytic site E53 out of a total of 68 ESTs covering that codon (BX258723, BX258724, BX261328, BX261329; Figure 3.4). All of these had the T/G change to create the TAG stop codon at bases 161, 173, 153 and 153, respectively, in their sequences. Nine other SNPs in ESTs were observed at this gene, including seven at nonsynonymous sites – four of these were non-conservative (Table 3.7). Resequencing of the lysozyme gene was conducted for a set of chickens as well as JF (grey, Ceylon, red, green) and related birds (bamboo partridge and grey francolin) in order to determine if these SNPs in ESTs were present in the extant population. See Chapter 7 for further details on amplification, SNP detection and sequence analysis of lysozyme. None of the EST SNPs detected at lysozyme were observed in the samples.

Table 3.7. EST SNPs identified at the lysozyme gene.

Position[1]	Exon	Ancestral	EST	Ancestral	EST	Nt position[2]	SNP type
51	2	K	N	AAA	AAT	1404	nonsynonymous
53[3]	2	E	Stop[4]	GAG	TAG	1408	nonsynonymous
63	2	R	A[4]	CGT	GCT	1439/40	nonsynonymous
64	2	N	R	AAC	CGA	1442/3/4	nonsynonymous
65	2	T	A	ACC	GCC	1445	nonsynonymous
70[3]	2	D	D	GAC	GAT	1461	synonymous
84	2	D	Y[4]	GAT	TAT	1502	nonsynonymous
85	2	G	C[4]	GGC	TGC	1505	nonsynonymous
102	3	L	L	CTG	TTG	3326	synonymous
108	3	S	S	GCG	GCT	3346	synonymous

[1] Amino acid codon. [2] Nt position is the GenBank Refseq nucleotide position affected. [3] Catalytic sites. [4] Non-conservative substitutions. None of these SNPs were detected in population resequencing of the gene.

Figure 3.6. Screenshot of Jalview alignment window for the TLR1LA gene.



```
                                      1440      1450      1460      1470      1480      1490
NM_001007488////1-236  TGAGCTGCACTCATTGCAGGAGCTGAACCTGGCATCCAACAGGCTGGCCGACCTGCCAGGGTGCAGGGC
20217918////1-563
20262326////1-349
20263374////1-338
25755644////1-2300     ---------------------------------------------------------------------
25844395////1-2242     ---------------------------------------------------------------------
25891052////1-2107     TGAGCTGCACTCATTGCAGGAGCTGAACCTGGCATCCAACCGGCTGGCCGACCTGCCAGGGTGCAGGGC
25910744////1-2364     ---------------------------------------------------------------------
25951071////1-2306     ---------------------------------------------------------------------
25965301////1-881
25965501////1-2195     TGAGCTGCACTCATTGCAGGAGCTGAACCTGGCATCCAACCGGCTGGCCGACCTGCCAGGGTGCAGGGC
28445557////1-855
53889094////1-938
53889095////1-1291
53889096////1-2062     TGAGCTGCACTCATTGCAGGAGCTGAACCTGGCATCCAACAGGCTGGCCGACCTGCCAGGGTGCAGGGC
53889097////1-2366     ---------------------------------------------------------------------
53889098////1-2366     ---------------------------------------------------------------------
53889099////1-2366     ---------------------------------------------------------------------
53889100////1-2364     ---------------------------------------------------------------------
53889101////1-2022     TGAGCTGCACTCATTGCAGGAGCTGAACCTGGCATCCAACAGGCTGGCCGACCTGCCAGGGTGCAGGGC
53889102////1-1355
53889103////1-773
53889105////1-122
56079401////1-2048     ---------------------------------------------------------------------
59952117////1-2364     ---------------------------------------------------------------------

Conservation           ---------------------------------------------------------------------

Quality

Consensus              TGAGCTGCACTCATTGCAGGAGCTGAACCTGGCATCCAACAGGCTGGCCGACCTGCCAGGGTGCAGGGC
```

The uppermost sequence shown is the genome reference sequence (NM_001007488.2). Amino acid 491 has a synonymous C to A SNP present in resequenced chickens at base 1471 (up from the red box). Dashes represent EST regions with no determined sequence. Sequence quality and consensus increase correspondingly with bar height.

3.4 Discussion

In order to identify polymorphisms in the chicken genome, ESTs were aligned to chicken genes and SNPs were detected according to conservative criteria. These were assembled into a local searchable interactive database that displayed the sequence alignments. Genes related to immunity and SNPs causing stop codons were investigated in particular because of their importance in relation to disease.

3.4.1 EST database development:

The construction of a searchable database was an effective strategy for managing a large EST dataset. This asset provided a framework for further analysis of resequenced genes in the event that such work discovers the same SNPs as those identified by analysing ESTs. This could deliver an efficient approach for determining valid ESTs in the chicken, as demonstrated by the observation of a SNP present in ESTs in the broiler population at TLR1LA.

Although this synonymous EST mutation at TLR1LA was not functional, it was segregating at an intermediate frequency in the commercial broiler population but was not observed in the small red JF population sample. It was present in two sequences from bursal lymphocytes of Prague CB chickens (Caldwell et al. 2005) and in one of the BBSRC chicken sets (Boardman et al. 2002) but was not observed in the red JF genome reference sequence (International Chicken Genome Sequencing Consortium 2004). Thus this mutation could have emerged since domestication, however only comprehensive resequencing in divergent red JF populations could confirm this.

3.4.2 EST SNP discovery:

The limited number of chicken protein-coding genes in this analysis (3,154) was indicative of the restricted set of ESTs sequenced by previous EST-generating studies. The observations of more synonymous than nonsynonymous and more conservative than non-conservative substitutions illustrated the conserved nature of this dataset. The consistent emergence of pancreatic amylase (AMY2A), haemoglobin gene (HBA1) and lactate dehydrogenase B (LDHB) as highly polymorphic genes that had known functions suggests these may be good candidates for further investigation of SNPs in ESTs. Additionally, the immune gene CD3ε, which spans the cell membrane as the signal transducing element of T cell antigen receptors, has undergone adaptation since the avian-mammalian divergence and hence also represents a gene of interest (Göbel & Fluri 1997, Gouaillard et al. 2001).

Further analysis of genes implicated in immunity suggested a further set of 19 genes that may be subject to selection due to their high level of nonsynonymous or non-conservative changes ($P_N/P_S \geq 0.5$ or non-con/con ≥ 0.5), particularly at known genes IRAK2, IL16 and TIMD4. The latter has a high incidence of nonsynonymous changes that may be a reflection of alternate splicing, which is present in humans (Park et al. 2009), and would increase the detection of mutant variants.

An examination of stop SNPs at immune genes revealed one candidate gene, lysozyme, that was resequenced because of the significance of a candidate EST stop SNP and the crucial role of the gene in innate immune defence (Holler et al. 1975a, Holler et al. 1975b). Although the samples resequenced were diverse and included bird species related to chicken, none of the SNPs were observed. This has important implications for the approach for detecting SNPs in ESTs implemented here, however, these must be tempered by detection of one EST SNP at the TLR1LA gene.

Previously EST SNP analyses in other organisms have been successful: half of such substitutions were detected in 61 resequenced cattle genes (Hawken et al. 2004). This approach was developed further, so that many novel SNPs in bovine ESTs could be identified computationally and verified through resequencing (Lee et al. 2006). Thus the question arising here is why the chicken EST SNPs detected were not observed in subsequent resequencing.

Many SNPs in ESTs may be sequencing errors, caused in part by the unknown sequence quality of ESTs (Hawken et al. 2004). One response would be to increase threshold requirement beyond four ESTs – however, the TLR1LA EST SNP and ancestral allele were detected in just three sequences each. Such quantification is complicated by the possibility of EST libraries sequencing the same individual for the same gene more than once (Hayes et al. 2007). A further modification that could improve power to detect SNPs is to use more stringent criteria for the clustering steps (Wang et al. 2004), though those used here are arguably conservative, given the short lengths of most ESTs. These results may be a reflection of the low concentration of expressed immune genes in the tissues samples to create the EST libraries.

A SNP-based analysis has suggested that commercial chickens have a 50% deficit of rare alleles (Muir et al. 2008), and given that the effectiveness of SNP-based approaches may be reduced by ascertainment bias leading to the absence of low-frequency variants, the lower SNP site variation present in this study suggests that ESTs may have less diversity as well. Nonetheless, the chicken's geographically and genetically admixed history suggests this may not be a dominant feature of the chickens resequenced (Kanginakudru et al. 2008). Additionally, the absence of a population bottleneck during chicken domestication (International Chicken Genome Sequencing Consortium 2004) compared to other domestic species, such as the cow (Finlay et al. 2007), would be expected to have maintained more nucleotide variation. Moreover, chickens do display high levels of diversity, even among commercial birds (International Chicken Genome Sequencing Consortium 2004).

The resequenced lysozyme gene displayed extensive CDS conservation: only three SNPs were observed, and two of these were singletons. TLR1LA had 24 coding SNPs, but 18 of these were singletons. Hence it is possible chicken ESTs will only detect intermediate- and high-frequency SNPs, such as that observed at TLR1LA, which was polymorphic in half of the broiler genotypes.

While none of these explanations alone are sufficient to explain the lack of power present in this analysis, when taken with the unknown nature of EST library sequence quality, they serve as a basis for developing better analyses of diversity in domestic species, particularly at the advent of high-throughput transciptome resequencing (for example, Trick et al. 2009, Ng et al. 2009).

Acknowledgements

Three perl scripts (run_analysis3.pl, chick_nonSNPit.pl and run_analysis_nocap3.pl) and Figure 3.1 were modified from versions originally created by David Lynn (Simon Fraser University, Canada).

Publication

Part of the methods used in this chapter formed the basis for generating a portion of the zebra finch DNA sequence dataset in a publication in *Developmental & Comparative Immunology* 33(9):967-73 in 2009 entitled “The Avian Toll-Like-Receptor Pathway – subtle differences amidst general conformity”. The authors are: Cormican P, Lloyd AT, Downing T, Connell S, Bradley DG and O’Farrelly C.

CHAPTER 4

Contrasting evolution of diversity at two disease-associated chicken genes

4.1 Introduction

4.1.1 Interleukin and interferon as key cytokine families:

Cytokines are communicating proteins that modulate the effects of the immune response by binding target cell receptors (Charo & Ransohoff 2006). Analysis of protein structural homology has differentiated this group of molecules into several families: the interferon and interleukin gene families encode cytokines that comprise a substantial portion of the immune genes in the chicken genome (Kaiser 2007). Chicken interleukins are signalling molecules that are secreted by immune cells to regulate and activate the immune response (Kaiser 2007). Human interferons can inhibit the replication of viruses, upregulate lymphocyte antigen presentation and activate natural killer cells as well as macrophages (Sen 2001).

Chicken interferons and interleukins have been categorised into families based on synteny with their human orthologs (Table 4.1). It is likely that continued genome annotation and analysis will reveal more interferons and interleukins (such as those in Fumagalli et al. 2009), many of which may be novel variants that are still unidentified due to the use of a mammalian comparison. The use of avian comparisons, such as the chicken and zebra finch, has been effective for other studies seeking to discover immune genes (for example, Cormican et al. 2009). In addition, the general chicken genomic pattern of gene family size reduction and gene loss reinforces the likelihood that the remaining members are not redundant and are functionally pivotal to chicken immune defences.Interleukin 1-β (IL1B) and IFNG are central components of the immune response with distinct patterns of function. These two genes were selected here for resequencing in chicken populations.

Table 4.1. Chicken cytokine family types and sizes.

Group	Family	Number
Interleukin[1]	T-cell proliferative	3
	IL1	1
	IL10	4
	T_H1	3
	T_H2	6
Inteferon[2]	Type I	11
	Type II	1

Of the 26 known interleukin genes, five remain unassigned (Kaiser 2007, Kaiser et al. 2005). Of the 12 known interferon genes the sole type II one is interferon-γ (IFNG; Kaiser 2007, Kaiser et al. 2005).

4.1.2 The role of IL1B and IFNG in immunity:

The sole known type II chicken interferon, IFNG, is a central mediator of immune networks (Kogut et al. 2005a). The human form activates the Jak-Stat (Janus kinases – signal transducers and activators of transcription) pathway by binding a heterodimer consisting of IFNG receptors -1 and -2 (Hebenstreit et al. 2005). IFNG has key roles in innate and adaptive immunity, being an important product of T_H1 lymphocytes (Schoenborn et al. 2007) and can inhibit viral replication: for example, in infectious bursal disease in chickens (Rauw et al. 2007). Predominantly expressed by natural killer cells and natural killer T cells, the main role of human IFNG is to activate and mediate killing mechanisms including T and natural killer cell cytotoxicity (Schoenborn et al. 2007).

The chicken IL1B gene is a member of the IL1 family and expresses a pleiotropic proinflammatory cytokine that activates cells by binding the IL1R, and is thus involved in innate immune responses to a wide variety of signals (Weining et al. 1998, Kaiser et al. 2005). This protein is produced in significant amounts during the early stages of inflammation, generally after the activation of TLR signalling (De Nardo et al. 2005). Important secretors of human IL1 include macrophages, dendritic cells and epithelial cells. It is expressed as a proprotein before being converted to its active form by caspase-1 (also known as ICE, interleukin-1 converting enzyme), which in human can be activated by certain lipoproteins, implicating IL1B in causing inflammation associated with the thickening of artery walls (Stollenwerk et al. 2005). IL1 has additional roles beyond the immune response, including regulation of the cell cycle (Madge et al. 2000), cranial hormone regulation (Spangelo et al. 2000) and apoptosis (Bratt and Palmblad 1997).

Human interleukins may regulate expression of the human Mx gene (Simon et al. 1991). Similarly, mouse IL1 and IFNG alter the expression of the mouse Mx gene (Goetschy et al. 1989). Balancing selection may be present in the 5' UTR of human IL1, and high diversity has been observed at the 5' end of IFNG and IL1 (Hughes et al. 2005).

4.1.3 IL1B and IFNG and chicken disease:

The chicken IFNG gene is known to affect resistance or susceptibility to several diseases. It is implicated in susceptibility to infection with *Salmonella* (Ye et al. 2006) – IFNG expression levels are low in chickens susceptible to *SE* serovar enteritidis, while resistant birds tend to express it at a higher level (Sadeyen et al. 2004). Interestingly, IFNG appears to mediate transcription and expression of other cytokine genes (including IL1B) in response to *SE* serovar enteritidis infection (Kogut et al. 2005a). IFNG also enhances the immune response to *Escherichia coli* infection (Janardhana et al. 2007) and to *Brucella abortus* antigens (Zhou et al. 2001). In the duck, IFNG expression has protective qualities against both duck hepatitis B virus and vesicular stomatitis virus when used in tandem with a vaccine (Long et al. 2005).

IL1B is implicated in the immune response to many pathogens, some of which have been documented. Like IFNG, IL1B responds to *Salmonella* infection: Okamura et al. (2004) found increased expression of IL1B following *SE* challenge. *Campylobacter jejuni* induces the expression of cytokines, including IL1B (Smith et al. 2005). IL1B expression responds to lipopolysaccharides and peptidoglycan (Kogut et al. 2006), and to recombinant human lipopolysaccharide-binding protein (Kogut et al. 2005b).

This analysis was an initial investigation into the evolution of specific functional immune genes in a range of diverse global chicken populations and related *Phasianidae* family species. Population-wide diversity present at IL1B and IFNG was complex and evidenced the diverse origins of the domestic chicken. The genes possessed distinct genetic characteristics, reflecting their functional roles and therefore indicating differing pressures have shaped their evolution. Comparisons with other *Gallus* species suggest red JF was the main origin of these genes in the domestic chicken.

4.2 Methods

4.2.1 Sample collection:

A total of 70 chicken samples were acquired from each of 3 Asian populations (from Bangladesh, Pakistan and Sri Lanka) and 4 African populations (from Botswana, Burkina Faso, Kenya and Senegal). Village chickens represent a reservoir of diversity, which may be useful in breeding in light of the reduced variability in certain commercial lines (International Chicken Genome Sequencing Consortium 2004). Given that chickens were domesticated in Asia, it is anticipated that much genetic diversity is preserved in these populations. The African samples were acquired for comparative purposes and to give a more global illustration of gene diversity in these two continents.

Three outgroup samples were obtained from the Department of Ornithology and Mammology at the Californian Academy of Sciences (CAS). The samples were green JF (CAS number 85707), Chinese bamboo partridge (CAS number 89821) and grey francolin (CAS number 87894). These were the same samples studied by Kaiser et al. (2007), who showed that green JF, bamboo partridge and grey francolin are the most closely related bird species to chicken, in that order (excluding grey, red and Ceylon JF). One grey, Ceylon and red JF each from Wallslough Farm (Co. Kilkenny, Ireland) were also sampled, giving a total of 6 outgroup samples. The DNA was isolated from the samples using phenol-chloroform extraction following proteinase K digestion.

4.2.2 Sequence determination and acquisition:

Potential transcription factor binding sites at IL1B and IFNG were ascertained by orthologous alignments of these genes in chicken, human and mouse using Mulan (http://mulan.dcode.org). The UCSC (http://genome.ucsc.edu), GenBank and Ensembl (http://www.ensembl.org) genome browsers were used to map the structures of the genes. PCR primers were constructed using Primer3 software (http://frodo.wi.mit.edu) subject to having lengths of 20-24 bases, a GC content of about 50%, an annealing temperature of approximately 60°C, and a GC clamp. They were created by VHBio (UK, www.vhbio.com). The details of the actual primer pair sequences and optimal parameters for their usage were determined (Table 4.2). Each amplicon was amplified according to a PCR cycle setup (Table 4.3).

Table 4.2. Sets of primer pair sequences used in the amplification of IL1B and IFNG and their associated optimal PCR parameters.

Gene	Primer	Size (bp)	Orientation	T_M (°C)	[MgCl] (mM)	Primer Sequences
IL1B	1	650	Forward	59	20	CTTCACCCTCAGCTTTCACG
			Reverse			CTTCTGGTTGATGTCGAAGATG
	2	712	Forward	61	15	CTTCGACATCTTCGACATCAAC
			Reverse			ATACGAGATGGAAACCAGCAAC
	3	737	Forward	59	15	TCATCTTCTACCGCCTGGAC
			Reverse			CGCATTCGTTTGTGTAAGAAAG
	4	772	Forward	60	15	TCAGAGCCCTCTATCACTCCTC
			Reverse			CATCACGTAAACACTCGCTCTC
IFNG	1	728	Forward	61	20	GTCTAGTACCCACCCTGCATTC
			Reverse			GAAGTTTCTTTTACCCGTGGTG
	2	875	Forward	60	20	ACCAGAAATGAGTTGACTGTTG
			Reverse			CTGCGTTAAGAGCCACTGTATG
	3	687	Forward	61	20	CTTCAGCTGGGATTAGTCATACAG
			Reverse			GGGTCAGAGTTTAACCATCAGG
	4	822	Forward	61	20	GCTGACGGTGGACCTATTATTG
			Reverse			CCCAACTTCTAATCACCTGGAG
	5	663	Forward	60	20	CTGGAAAGTGTGATGTTTCCAC
			Reverse			GGAGGTCATAAGACGCCATTAG
	6	803	Forward	59	20	CCAGAATCTCTGTGAAAAGCAG
			Reverse			TCATTGTCTCACTGTTGGTTCC

Regions between primer pair numbers 3 and 4, and 5 and 6 of IFNG and upstream of pair 1 of IL1B were not successfully amplified.

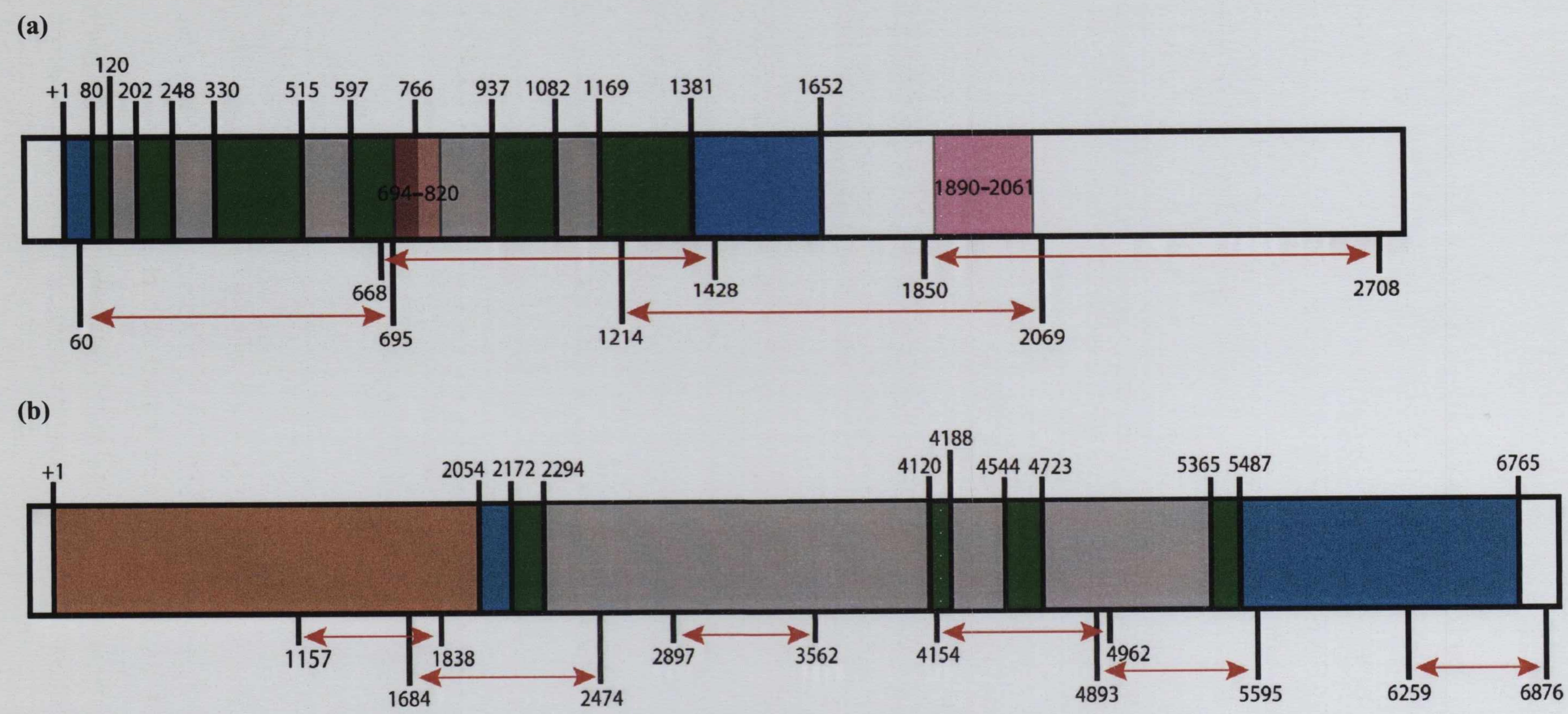
Table 4.3. PCR cycle program used for each primer pair.

Step	Temp. (°C)	Duration
1	95	15 mins
2	95	0.5 min
3	T_M [1]	0.75 min
4	72	1 min
5	72	15 mins

[1] Annealing temperature as listed in Table 4.1. Steps 2 to 4 were repeated 33 times in sequence.

Four amplicons covering IL1B and six for IFNG were successfully amplified by PCR for all 70 chicken and 6 outgroup samples (Figure 4.1). The forward and reverse PCR product sequences were determined by Agowa, Germany (http://www.agowa.de). The 5' end of the IL1B gene was not successfully amplified, a problem also encountered by Kaiser et al. (2004, although they did amplify a part). Consequently, the IL1B promoter and upstream regions remain unexplored here.

Figure 4.1. Gene structures of (a) IL1B and (b) IFNG.



Exons are shown in green, introns in grey, intergenic regions in white, UTRs in blue, amplicon regions by the red arrows and the promoter sequence is beige. The areas in light pink are those where the sequence quality was poor and were not included in analysis. The numbers shown represent the base positions in relation to the GenBank gene transcription start site. The genes differ in length.

4.2.3 Sequence assembly:

The DNA base sequencing completed by Agowa generated chromatograms that were assembled into contigs using the Phred-Phrap-Consed-Polyphred pipeline (http://www.phrap.org/phredphrapconsed.html) programs Phrap v0.990319 and Phred v0.020425.c (Ewing and Green 1998, Ewing et al. 1998). This pipeline has been used widely to detect SNP data, for example in chicken ESTs (Kim et al. 2003). After running the program Sudophred, the stringency used for clustering the contigs in Phred-Phrap was modulated by the forcelevel flag, which was set to a value of 10, allowing the maximum numbers of sequences to be incorporated into the same assembly.

Bases were called using Consed (Gordon et al. 1998): each SNP suggested by Phrap was verified independently and separately by two individuals. Using the formula $P(base\ is\ correct) = 1 - 10^{-S/10}$, where S the base quality score, most bases have a 99.99% or greater probability of being valid ($S \geq 40$; Johnson and Slatkin 2005). Any bases with a quality score of 14 or less were not included in the analysis, so that all bases had at the lowest a 96.8% probability of being correct. Similarly, only SNPs with high probability of being accurate (in polyphred ranks 1, 2 or 3) were selected for further examination. Polyphred version 5.0 (Stephens et al. 2006, Nickerson et al. 1997) was used to assemble the data for further processing using a series of Perl scripts.

The reference coding sequences were aligned against complete region sequences so that relative exon positions were confirmed by MGalign version 3.1 (Lee et al. 2003) and a list of the genotypes for each sample was collated.

Perl scripts (ppoutParser93.pl in Appendix A) were used to remove any sequence sites where there was inadequate coverage in populations, sub-standard base quality scores, or insufficient coverage for either forward or reverse sequences. PHASE version 2.1.1 (Stephens et al. 2001) was used to reconstruct the haplotypes and to infer any missing ones for 100 iterations with a burn-in of 100 iterations. A series of Perl scripts (PhaseIn.pl, SeqBuild.pl, hashadd.pl and maskseqs2.pl) as well as a Microsoft Excel application were used to parse the genotypes for use in PHASE, and so that the sequences could be exported to Mega (version 4.0.2; Tamura et al. 2007). There, the sequences were converted to formats for other software packages. A total of 2,351 bp of IL1B and 4,066 bp of IFNG were used for further analysis.

Haplotypes were assigned using PHASE and these were cross-referenced with haplotypes generated by Arlequin version 2.001 (Schneider et al. 2000) to ensure consistency: the haplotypes generated by both were identical. The nucleotide sequences were submitted to the GenBank: the accession numbers are FJ537713 to FJ537864 for IL1B and FJ537865 to FJ538016 for IFNG.

4.2.4 Generating summary statistics:

Nucleotide diversity was measured using π, the average number of nucleotide differences per nucleotide site between each sequence pair in a population: this is a basic measure of heterozygosity (Tajima 1983). This was calculated using DnaSP 4.0 (Rozas et al. 1999, Rozas et al. 2003). For a diploid neutral population of effective population size N_e and with mutation rate per locus per generation μ, the expected value of π is $4N_e\mu$ with variance $4N_e\mu(4N_e\mu + 1)$ – departures from this value may indicate non-neutral evolutionary or demographic effects.

Different demographic histories were modelled for average pairwise nucleotide difference (π) frequencies using DnaSP. Timing of ancient population size changes were estimated for the peak pairwise differences such that:

$$t = \frac{\tau}{2\mu m_T} \qquad \text{Equation 1}$$

where t is the estimated time of the event; μ is a mutation rate of 1.8×10^{-9} substitutions per site per year (Axellson et al. 2005); τ is the mean pairwise difference value – analogous to the time of the mean peak change in population size; and m_T is the gene length (Rogers & Harpending 1992). One generation was calibrated as one year.

DnaSP was used to analyse the polymorphic characteristics of the data and to perform a series of population genetic tests. The numbers and types of SNPs were assessed, as was the GC content and the number of alleles. The haplotype diversity (*Hd*) is related to the number (*k*) and distribution of haplotypes in the sample such that:

$$Hd = 1 - \sum_{i=1}^{k} p_i^2 \qquad \text{Equation 2}$$

where p_i is the frequency of the i^{th} haplotype, assuming an infinite-sites model (Depaulis and Veuille 1998). Thus the scale of *Hd* is from 0 (no haplotype diversity, Hd = $2(n-1)n^2$) up to 1 (maximal *Hd* of $(1-1/n)$) for a sample size *n*. This assumes no recombination – chicken genes with recombination therefore have higher *Hd* values than expected if this is present at an elevated level (Depaulis et al. 2001).

Kelly's Z_{nS} is a measure of LD based on haplotype correlations between sites (Kelly 1997). For the allele frequencies of sites *i* and *j*, p_i and p_j, the LD (denoted D_{ij}) between each is: $D_{ij} = p_{ij} - p_i p_j$, where p_{ij} is the frequency of sequences with derived alleles at both sites. A standardised form of this value is δ_{ij}, ranging between 0 and 1: $\delta_{ij} = \dfrac{D_{ij}^2}{p_i p_j (1-p_i)(1-p_j)}$. From this the statistic Z_{nS} is taken as the average δ_{ij} for all sequence pairs with *S* segregating sites:

$$Z_{nS} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} \delta_{ij} \qquad \text{Equation 3}$$

where *n* is the number of sequences sampled. Recombination is likely to disturb gene-wide patterns of LD by placing neutral sites adjacent to those under selection, and by generating new variants at pairs of sites (Kelly 1997).

Watterson's coalescent estimator of variation, θ_W, assumes no population structure, an infinite-sites model and a large N_e compared to the number of genotypes sampled (*n*) (Watterson 1975). For the number of SNPs per nucleotide (*S*), θ_W is determined as:

$$\theta_W = S / \sum_{i=1}^{n-1} i^{-1} \qquad \text{Equation 4}$$

This statistic measures the abundance of rare alleles in the sample and can estimate $\theta_W = 4N_e\mu$ for a diploid population.

Recombination is a combination of processes that occurs during meiosis where homologous DNA regions cross-over to exchange sequence or non-reciprocally convert sequence of other homologous regions (Betran et al. 1997). Recombination was studied using the four gamete test to obtain the minimum number of recombination events, R_M – however, this statistic may substantially underestimate the actual number events that have taken place (Hudson and Kaplan 1985). The expected value of this statistic is dependent on the number of samples (n) and the rate of recombination per site per generation (c):

$$E[R_M] = 4N_e c / \sum_{i=1}^{n-1} i^{-1} \qquad \text{Equation 5}$$

However, the actual algorithm implemented here with DnaSP to measure R_M (using DnaSP) is more complex (see Appendix 2 in Hudson and Kaplan 1985). Hudson and Kaplan regard R_M as analogous to the number of observed SNPs in Equation 4 above: it measures observed recombination but cannot be use to infer total recombination, just as SNP sites are observed variation but cannot predict the true total number of variable sites. R_M can be visually apparent in phylogenetic networks, where unresolved branching between sequences is reflected by the value of R_M: so as R_M increases, so does the number of unresolved phylogenetic branch trifurcations.

The rate of recombination, R, is based on c, the recombination rate per site per generation:

$$R = 4N_e c \qquad \text{Equation 6}$$

assuming an infinite-sites model (Hudson 1987). The variance of R is calculated in the same way as π stated above: $4N_e c(4N_e c + 1)$. This is based on Watterson's θ_W – a metric to which R is analogous (Hudson and Kaplan 1985). Following from Equation 6, the estimated value of R is then adjusted for the average distance between sites that have recombined. Because R is dependent on the inferred effective population size, it can be biased by the effects of both selection and demographic changes: directional and diversifying selection or a population size increase cause elevated N_e values, and purifying selection or a population bottleneck can yield reduced N_e values.

Understanding the extent of gene conversion in resequenced data is important because of its effect on allele diversity. The number (N) and lengths of gene conversion tracts occurring pairwise between populations that possess sufficiently divergent haplotypes are dependent on the probability of detecting a converted site (ψ) and the average true tract length ($(1 - \Phi)^{-1}$) with variance ($(1 - \Phi)^{-2}$), where Φ is a geometric distribution parameter of the true tract lengths, which differ from the observed tract lengths (Betran et al. 1997). This gives the probability that a single gene conversion event creates a tract of length m equal to $(1 - \Phi)\ \Phi^{m-1}$, and to detect such a tract at least two symptomatic SNPs are required to be observed so the chances of detection are $1 - (1 - \psi)^m(1 - m\psi/(1 - \psi))$. From this it is possible to deduce the difference between the observed (x) and true mean tract lengths through ψ, where $x = d + 1 - g$ for d, the distance between the most remote affected bases, and g, the number of nucleotide gaps between these bases, and ψ in this instance is estimated as average probability of detecting a set of sites in a specific tract. The number of undetected gene conversion events and the rate of gene conversion per generation can also be inferred (see Betran et al. 1997 for details).

4.2.5 Assessing population structure:

F_{ST} is a genetic fixation index of relative population differentiation that takes π within a population (π_p) and the mean π between the initial population and another (π_d) such that $F_{ST} = 1 - \pi_p/\pi_d$ (Wright 1951). Thus F_{ST} is a measure of relative heterozygosity between populations that is scaled from 0 (no difference) to 1 (maximal difference), though these are relative rather than absolute measures.

In order to further examine possible links between genetic differentiation and population configuration, pairwise F_{ST} values were calculated by Arlequin for each population. The most parsimonious population tree structure was determined by Neighbor from Phylip version 3.67 (Feselstein 1995). Treeview (Page 1996) was used to produce trees of the populations. Using PASSaGE (Rosenberg 2007), Mantel permutation two-tailed tests (Mantel 1967) were performed for 10^7 permutations between the pairwise F_{ST} values for each gene and the geographic great circle distances between the samples' countries.

Analysis of Molecular Variance (AMOVA) tests (Excoffier et al. 1992) were conducted on all sites using Arlequin, with 1,000 permutations. This assigns the observed variation to different components of population structure using an ANOVA (Analysis of Variance) approach: within populations, between populations and between continental groups. This test assumes panmictic populations and non-random mating. It is an effective test of the population structure present at these genes because of the global sampling conducted.

4.2.6 Creating haplotype networks:

Phylogenetic allele networks are an effective means of communicating the genetic relationship between genotypes. The large number of samples under examination here requires effective methods for clustering haplotypes because there are many possible configurations. Mutations that are reversals of previous states or replicate other non-ancestral states complicate genetic networks, and result in a computationally large number of possibilities (Bandelt et al. 1995). The method used here created median-joining networks, which work by firstly making a set of minimally spanning trees from the genotypes sequences separated by mutations (Bandelt et al. 1999). It secondly inserts intermediate (possibly ancestral) nodes where there are shared mutations between at least three adjacent genotypes, shortening the overall distances. Finally, this approach optimises the trees using a maximum parsimony heuristic algorithm to produce the network with the shortest total mutational distance for all sequences, which can include unresolved ancestries between haplotypes (Bandelt et al. 1999) – these could be recombination events and can be estimated using R_M in Equation 5. Median-joining networks perform more accurately than those relying on using the minimum spanning methodology alone (Woolley et al. 2008).

The median-joining haplotype networks were constructed for IL1B and IFNG using Network version 4.2.0.1 software (Bandelt et al. 1999; http://www.fluxus-technology.com). Networks in this thesis were designed without pre- or post-processing steps and with the criteria for joining nodes as the connection cost rather than a greedy algorithm (greedy FHP), which joins the nearest nodes at each iteration. However, the greedy FHP was used for GMCSF (Chapter 7) because the connection cost criteria did not converge on a single network. Extensive recombination that was inferred at certain genes was in part represented by the nodal points in the networks, and this recombination led to the presence of certain mutations more than once in the network.

Phylogenetic networks can yield clues regarding the recent history of population genotypes: a star-like radiation of closely related samples may indicate positive selection (Bamshad & Wooding 2003). Clustered nodes may signify balancing selection, or other effects like selective pressures in local environments. The networks produced here might be best viewed as a representation of the distribution of extant diversity within chicken, rather than an accurate diagram of the genes' or samples' ancestry.

4.2.7 Predicting impact of amino acid replacement mutations:

Predictions to estimate the extent of the functional impact of each nonsynonymous substitution were conducted using SIFT (Ng & Henikoff 2003), PMut (http://mmb2.pcb.ub.es:8080/PMut/; Ferrer-Costa et al. 2005) and PolyPhen (http://genetics.bwh.harvard.edu.pph/; Ramensky et al. 2002). These operate by aligning the protein sequence of interest to available sequence homologs and determining the probabilities of particular replacement changes based on this data as well as the physical properties of the amino acids involved. In many cases the results of estimating the extent of functional impact for each nonsynonymous substitution were not statistically supported or had conflicting results, most likely due to the high protein sequence divergence between chicken and the species to which it was compared. Thus the mutation outcomes were also classed as not determined or probably neutral, depending on the output of the programs.

4.2.8 Tests of neutrality:

Summary statistics were used to evaluate the degree of deviation from neutrality: Fu and Li's *D* and *F* (Fu and Li 1993), Tajima's *D* (Tajima 1989), Fu's F_s (Fu 1997), Fay and Wu's *H* (Fay and Wu 2000). This set of statistics are based on examining the relative ratios of singleton, rare, intermediate and high frequency alleles in the populations, whose proportions should be approximately equivalent under neutral conditions. These statistics were determined using DnaSP using an infinite-sites model based on the number of mutations; the number of segregating sites can also be used to calculate their values, however this alternate method gave the same values for each metric.

The first published of these, Tajima's *D*, compares moderate frequency alleles (π) and the relative number of segregating sites (*S*), which reflects the number of rare variants more strongly. These statistics are adjusted for the sampling size according to:

$$D = \frac{\pi - S / a_1}{\sqrt{e_1 s + e_2 s(s-1)}} \qquad \text{Equation 7}$$

where $a_1 = \sum_{i=1}^{n-1} i^{-1}$ and $e_1 = c_1 / a_1$ such that $c_1 = b_1 - 1/a_1$ with $b_1 = \frac{n+1}{3(n-1)}$ and $e_2 = \frac{c_2}{a_1^2 + a_2}$ for $c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$ where $a_2 = \sum_{i=1}^{n-1} i^{-2}$ and $b_2 = \frac{2(n^2 + n + 3)}{9n(n-1)}$ (Tajima 1989). This test of neutrality is scaled such that the denominator is the standard deviation of the numerator, and so under ideal conditions behaves like a normal distribution ~ N(0,1). Under recombination, the variation of *D* decreases, so simulations were used to adjust for such effects that violate neutral assumptions. Tajima's *D* is robust to low numbers of samples and SNPs, and is an efficient approach for assessing neutrality in populations where ancestry is complex.

Additional tests of neutrality examine the relative difference between the total number of mutations (η), π and the number of derived singletons (η_e). η is inferred from an outgroup sequence to identify the number of singletons on internal branches (η_i), such that $\eta = \eta_e + \eta_i$ (Fu and Li 1993). The expectation of η_e is Watterson's estimator of genetic diversity from above, $E[\eta_e] = \theta_W = 4N_e\mu$, and $E[\eta_i] = \theta_W(a_n - 1)$ where $a_n = a_1$ above and n is the number of genotypes sampled. Fu and Li's *D* is the adjusted difference between η and η_e, testing the difference between the number of low frequency variants and the number of unique derived alleles. Fu and Li's *F* is the difference between π and η_e and so examines the ratio of alleles at intermediate frequency to the number of derived singletons (Fu and Li 1993). They operate in the same manner as Tajima's by adjusting for the variance between the statistics, while also incorporating sample size information. The test statistic for Fu and Li's *D* is:

$$D = \frac{\eta - \eta_e a_1}{\sqrt{u_D \eta + v_D \eta^2}}$$

Equation 8

where $b_n = a_2$ above and $v_D = 1 + \frac{a_n^2}{b_n + a_n^2}\left(c_n - \frac{n+1}{n-1}\right)$ with $c_n = \frac{2na_n - 4(n-1)}{(n-1)(n-2)}$, and $u_D = a_n - 1 - v_D$. The statistic for *F* is:

$$F = \frac{\pi_n - \eta_e}{\sqrt{u_F \eta + v_F \eta^2}}$$

Equation 9

where $v_F = \left[c_n + \frac{2(n^2+n+3)}{9n(n-1)} - \frac{2}{n-1}\right]/(a_n^2 + b_n)$, with n, b_n, c_n and$\pi_n = \pi$ as above, and $u_F = \left[1 + \frac{n+1}{3(n-1)} - 4\frac{(n+1).(a_{n+1} - (2n/(n+1)))}{(n-1)^2}\right]/a_n - v_F$. *D* and *F* are effective tests for the presence of background selection: in such a case a deficit of singletons would be expected – an excess may reflect a population expansion. They are sensitive to low numbers, and thus sufficient numbers of externally and internally branched singletons are required for confident testing. If ancestry is not determined and hence the direction of the mutations are unknown, Fu and Li's *D* and *F* can be calculated as *D** and *F**, respectively, however these metrics are less powerful and since outgroups were acquired, were not implemented here (Fu and Li 1993).

Fay and Wu's H examines the relative difference between variants at intermediate (θ_π) and high (θ_H) frequencies, where the latter is determined for derived alleles from outgroup information as $\theta_H = \sum_{i=1}^{n-1} \frac{2i^2 S_i}{n(n-1)}$ and $\theta_\pi = \sum_{i=1}^{n-1} \frac{2iS_i(n-i)}{n(n-1)}$, where n is as above, and S_i is the number of derived alleles found *i* times (Fay and Wu 2000). The *H* statistic is:

$$H = \theta_\pi - \theta_H \qquad \text{Equation 10}$$

H is effective even if the number of high-frequency variants is low. This metric can detect hitchhiking signatures brought about by positive selection because neutral alleles near positively selected sites are more likely to be fixed, a signal of hitchhiking can be detected in some cases (Bamshad & Wooding 2003).

Fu (1996) defined F_S as a measure of the relative abundance of rare alleles in a sample: a significant excess (F_S is highly negative) or deficit (F_S is very positive) can reflect selective and other processes. The statistic is:

$$F_S = \ln\left(\frac{S^l}{1 - S^l}\right) \qquad \text{Equation 11}$$

where $S^l = \sum_{k=k_0} \frac{|S_k| \theta_\pi^k}{S_n(\theta_\pi)}$ for a sample of k_0 alleles from a total of *k* alleles in *n* sequences, with the product $S_n(\theta_\pi) = \theta_\pi(\theta_\pi - 1)...(\theta_\pi - n + 1)$ and S_k is the coefficient of θ_{π}^{k} in S_n. Thus S' is equal to the probability of having k_0 or more alleles in a random sample, given neutrality ($\theta = \pi$). An excess of rare alleles detected by F_S can often be related to the actions of hitchhiking, population growth or recombination. Consequently, Fu's F_S is most useful when used in tandem with the above tests.

In order to test the neutrality of each summary statistic (including those mentioned in earlier sections), coalescent simulations of neutral data with the numbers of genotypes, segregating sites, total sites, sample population sizes and rate of recombination were performed for each gene using DnaSP for 1,000 replicates. These simulations generated empirical distributions with which the observed values were compared to determine the extent of their deviation from neutrality. Non-neutral evolution was inferred if the observed values lay at the extremes of the distribution. Sliding window analyses were performed for the summary and descriptive statistics, however, these were not informative for any of the genes resequenced in this thesis.

The HKA test can detect selection at loci that share common demographic history by seeking signatures of divergent evolutionary patterns between pairs of genes, where one shows particularly high levels of variability (Hudson et al. 1987). However, this test was not implemented in this thesis because of the hybrid nature of chicken domestication history; certain loci may not share the exact same genetic ancestry, which would ultimately produce false positive results. In addition, disparate domestication and introgression events would lead to further demographic effects on selection, such that comparing pairs of loci without context would not yield meaningful results.

4.2.9 Coalescent simulations to infer demography and recombination:

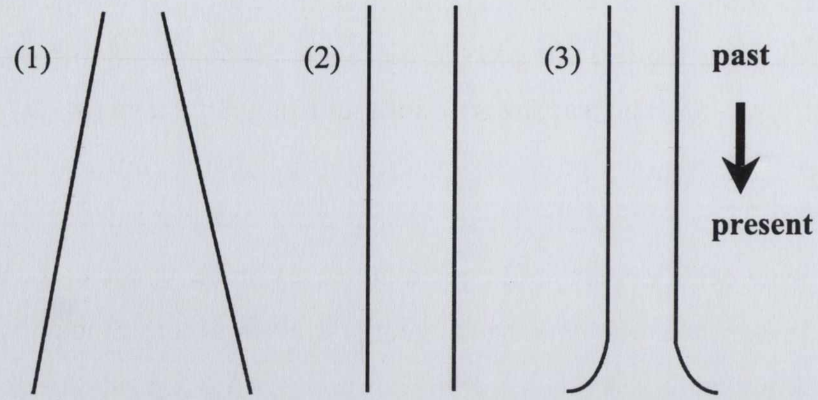
Recombination was estimated for the aligned gene sequences using the CLE (composite likelihood) method (Hudson 2001), which was applied with the program LDhat (McVean et al. 2002). This estimates the gene-wide rate of recombination, based on the number of samples and θ_W. Recombination is determined according to $\rho = 4N_e r$ per locus, constant across all sites. LDhat estimated recombination rates at the SNP loci, and tested locally elevated recombination at SNP hotspots. Initial simulations with LDhat normally assume $\rho < 100$, however IL1B exceeded this range and a limit of $\rho < 400$ was used.

Coalescent simulations in the form of samples under neutral models were generated for 10 repetitions using the program MS (Hudson 2002) and analysed with scanMS (Ardell 2004) using Perl scripts (parser_scanms.pl, Appendix A). The input data was generated by DnaSP (numbers of segregating sites, number of samples, cross-over rate, gene length). The degree of recombination at each simulated locus was calculated from the DnaSP results for *R*. The numbers of segregating sites were fixed and the simulations were completed both with and without migration. Tajima's *D* and π were simulated and the simulated values were compared to the observed DnaSP data to determine how demography affects the distribution of diversity at these loci. The strategy used here of examining neutrality in genes with given simulated recombination and demographic parameters has been implemented by others (Quesada et al. 2006, Ronald & Akey 2005).

Inter-population migration between seven populations was simulated within the constraints of MS and varied in order to ensure this reflected the admixture levels observed by AMOVA and the F_{ST} values. Migration has been previously used in MS models (Akey et al. 2004, Schaffner et al. 2005). The levels of migration were adjusted by optimising the parameter $4N_0m$, which determined the immigrant genotype composition of the subpopulations and the degree of migration between them. This permitted the examination of different demographic scenarios: the genes were simulated with demographic change and crossing over in order to test more carefully their neutrality. The distributions generated by scanMS (Ardell 2004) allowed analysis of the fit of the observed values on these models.

Different models of population history were designed to test the likelihoods of hypothetical demographic histories, according to three different scenarios (Figure 4.2). Models 1 (increasing) and 3 (exponential) had final population sizes (10^6) ten times greater than the initial population sizes (10^5). Model 2 (constant) was neutral and had a population of 10^6 throughout.

Figure 4.2. A representation of the demographic models simulated by MS.



(1) a gradual constant increase; (2) a constant population size; and (3) a static population until recent exponential population growth (to simulate a hypothetical domestication event).

4.2.10 PAML analysis of interspecies evolution:

One rigorous approach for examining selection pressure is to calculate the relative rate of nonsynonymous mutations (d_N) to the relative rate of synonymous mutations (d_S) in the protein-coding portion of a gene with $\omega = d_N/d_S$. Analysis of ω under different models was performed using the codeml implementation of PAML 3.15 package (see Chapter 2 for more information; Yang 1997).

The free-ratio (M1) model was used to calculate tree branch lengths and ω for each species lineage in the sample. For each model and each gene, chicken samples from the most frequent haplotype were used: one from Pakistan (FJ537719) for IL1B and one from Sri Lanka (FJ537935) for IFNG. Using different haplotypes yielded insignificant changes to results. The sequences used for all samples were all in the 5' to 3' orientation. The PAML models implemented are sensitive to low numbers of species, which totalled seven here (Anisimova et al. 2001). For all models and datasets, the presence or absence of gaps made no difference to the results of the alignments.

Lineage-specific models (M2) estimate one ω for one or more specified branches: the remaining branches have a different estimated ω (Yang 1997). This is then compared to a model with a fixed ω for all branches. For this and other codeml tests, a likelihood ratio test (LRT) was used to see if the estimated model is significantly more favoured than the neutral model according to a χ^2 distribution. The number of degrees of freedom is the number of parameters in the estimated model minus the number in the fixed model (Yang 1997). LRTs were conducted for lineages using two ω values (two-ratio models) and sets of lineages using more than two ω (multiple ratio models).

Site specific models (M1a and M2a, M7 and M8) estimate ω ratios for each site across the whole sequence by using a random sites model according to a Bayes empirical Bayes (BEB) model (Yang 1997, Nielsen and Yang 1998). For each model, ω and the proportion of sites affected were determined. For M1a only two ($K = 2$) fixed ω values are permitted: $\omega_0 = 0$ (conserved) and $\omega_1 = 1$ (neutral) with proportions $p_0 = 1 - p_1$. For M2a, these same two classes are allowed, along with an additional class where the ω ratio is freely estimated ($K = 3$) with proportion p_2 to allow for deviations from neutrality. A LRT is performed between the likelihoods of these two models. M7 is a neutral model that calculates $K = 4$ sites classes from a beta distribution, all of which are between 0 and 1. M8 has $K = 5$ with the same four beta-distributed classes as M7 with an additional class where $\omega > 1$. The LRT was calculated between M8 (the estimated model) and M7 (the fixed model), and similarly for M2a (the estimated model) and M1a (the fixed model), though the M8 versus M7 test is more effective. A BEB examination of the sites determines the posterior Bayesian probability of the ω ratio for each amino acid site. This differs from a fixed sites model because it uses a statistical model for ω variation, rather than using structural information, which can be challenging to determine for chicken genes. A significantly high posterior probability for this free ω class suggested a particular site is under positive selection, if M8 (or M2a) was significantly favoured and $\omega > 1$ (Nielsen & Yang 1998).

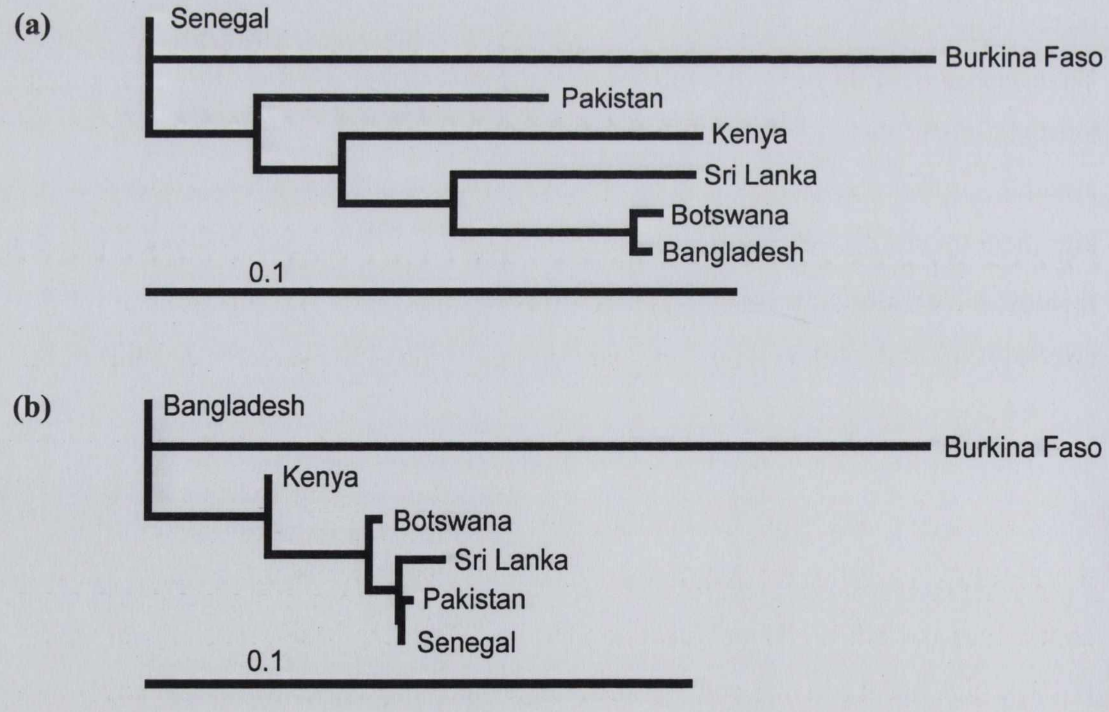
4.3 Results

The cytokines IL1B and IFNG are key regulators of chicken immunity and have been implicated in resistance to multiple chicken diseases. These genes were resequenced in seven chicken populations from Africa (Botswana, Burkina Faso, Kenya, Senegal) and Asia (Bangladesh, Pakistan, Sri Lanka) – a total of 70 samples for each gene. Additionally, four variants in the same genus as chicken were examined (red, grey, green and Ceylon JF) and two outgroup species closely related to chicken (bamboo partridge and grey francolin).

4.3.1 SNP and population diversity:

In order to determine if gene diversity was geographically structured, Mantel permutation tests were conducted the population pairwise F_{ST} values (Table 4.4) and neighbour-joining phylogenetic trees (Figure 4.3) were created from the F_{ST} values. Neither approaches showed evidence of an association between pairwise F_{ST} and geography at IL1B (Mantel p = 0.143) or IFNG (p = 0.143).

Figure 4.3. Neighbour-joining trees of populations sampled for (a) IL1B and (b) IFNG.



The scaled genetic distance shown is 0.1 substitutions per site.

Table 4.4. F_{ST} values for chicken populations for IL1B (top) and IFNG (middle) and geographic distances in miles (bottom).

IL1B F_{ST}	Pakistan	Burkina Faso	Senegal	Sri Lanka	Botswana	Bangladesh
Burkina Faso	0.218					
Senegal	0.027	0.112				
Sri Lanka	0.143	0.192	0.078			
Botswana	0.064	0.276	0.075	0.087		
Bangladesh	0.079	0.273	0.08	0.061	0.007	
Kenya	0.134	0.198	0.077	0.134	0.098	0.098
IFNG F_{ST}	Pakistan	Burkina Faso	Senegal	Sri Lanka	Botswana	Bangladesh
Burkina Faso	0.186					
Senegal	0.003	0.161				
Sri Lanka	0.009	0.206	0.010			
Botswana	0.003	0.188	0.014	0.015		
Bangladesh	0.029	0.126	0.045	0.024	0.029	
Kenya	0.026	0.141	0.006	0.015	0.001	0.002
Distances	Pakistan	Burkina Faso	Senegal	Sri Lanka	Botswana	Bangladesh
Burkina Faso	4,867					
Senegal	5,689	1,095				
Sri Lanka	1,887	5,589	6,644			
Botswana	5,075	3,147	3,995	4,287		
Bangladesh	1,207	5,989	6,886	1,366	5,492	
Kenya	3,374	2,797	3,885	3,078	1,764	4,009

The F_{ST} values at the top are those observed between populations at IL1B; those in the middle section are for IFNG. The greater circle distances between the countries in miles are shown in the section at the base.

The absence of strong population structure and abundance of variation was supported by AMOVA tests that assigned variation observed among all sites to different components of population structure. Most variation was found between individuals in populations – 87.5% of the variation at IL1B and 95.6% of that at IFNG was found at this level. Variation between populations accounted for 12.5% of diversity at IL1B and 4.4% at IFNG. Interestingly, no variation partitioned between the continents, Asia and Africa.

The frequency of SNPs at the genes was different: IL1B had 52 SNPs in 2,351 bp, whereas IFNG had just 68 in 4,066 bp (Table 4.5). The patterns of coding SNPs were contrasting as well: IFNG had just one SNP in 491 coding bases, much fewer than IL1B, which had 11 in 801 coding bases. High intra-population variation was observed in the numbers of SNPs in each continent. Two of nine nonsynonymous

SNPs in total at IL1B were segregating at moderate frequencies: 9.3% at base 952 and 10.7% at base 1261 (Table 4.6, Table 4.7) and their minor alleles occurred in both continents (Table 4.8, Table 4.9).

Table 4.5. Number of SNPs in the domestic chicken samples.

SNPs	Total	Informative	Singleton	Non-coding	Coding	
					Synonymous	Nonsynonymous
IL1B	52	50	2	41	2	9
IFNG	68	60	8	67	1	0

Table 4.6. IL1B coding SNP positions and types and the details of the observed amino acid substitutions associated with them.

Position	Type	Base change	Frame	Amino acid allele	
				Major	Derived
94	synonymous	CCG to CCA	3	Proline	-
225	nonsynonymous	GGC to AGC	1	Glycine	Serine
630	nonsynonymous	TCC to TTC	2	Serine	Phenylalanine
633	nonsynonymous	GCC to GTC	2	Alanine	Valine
939	nonsynonymous	GGG to AGG	1	Glycine	Arginine
952	nonsynonymous	GGG to GAG	2	Glycine	Glutamate
960	nonsynonymous	GGG to AGG	1	Glycine	Arginine
982	nonsynonymous	ACC to ATC	2	Threonine	Isoleucine
1063	nonsynonymous	GGG to GAG	2	Glycine	Glutamate
1181	synonymous	GTA to GTC	3	Valine	-
1261	nonsynonymous	TCC to TTC	2	Serine	Phenylalanine

Table 4.7. IL1B coding SNPs population frequencies and distributions.

Position	Exon	Continents	Population(s) of Origin [1]	DAF [2]
94	1	Africa	Bur, Bot, Ken	0.021
225	2	Both	Pak, Bot, Ban, Ken	0.036
630	4	Both	Pak, Bur, Bot, Ban, Ken	0.036
633	4	Asia	Pak	0.021
939	5	Africa	Bur, Bot, Ken	0.036
952	5	Both	Pak, Sen, Ban, Ken	0.093
960	5	Africa	Sen	0.014
982	5	Both	Bot, Ban	0.021
1063	5	Africa	Bot	0.014
1181	6	Asia	Pak, Ban	0.029
1261	6	Both	Pak, Bur, Sen, Ken	0.107

[1] Pak is Pakistan, Bur is Burkina Faso, Sen is Senegal, Sri is Sri Lanka, Bot is Botswana, Ban is Bangladesh, Ken is Kenya. [2] DAF is the derived allele frequency.

Table 4.8. Chicken and outgroup genotypes for IL1B at SNP sites polymorphic in the chicken samples.

The synonymous and nonsynonymous sites are signified by the letter "Y". Samples are from Pakistan (GenBank accession numbers FJ537713-FJ537732), Burkina Faso (FJ537733-FJ537752), Senegal (FJ537753-FJ537772), Sri Lanka (FJ537773-FJ537792), Botswana (FJ537793-FJ537812), Bangladesh (FJ537813-FJ537832), Kenya (FJ537833-FJ537852), bamboo partridge (FJ537853-FJ537854), grey francolin (FJ537855-FJ537856), green JF (FJ537857-FJ537858), grey JF (FJ537859, FJ537861), Ceylon JF (FJ537860, FJ537862) and red JF (FJ537863-FJ537864). Bases with nucleotide A are in green, C in blue, G in yellow and T in red.

Position	94	153	165	176	225	254	287	320	322	528	630	633	672	673	675	679	907	918	939	952	960	962	1063	1191	1261	1456	1522	1531	1554	1559	1561	1670	1674	1716	1734	2056	2065	2093	2109	2122	2129	2132	2151	2190	2269	2271	2272	2284	2357	2398	2403	2409
Synonymous	Y																								Y																											
Radical					Y						Y	Y							Y	Y	Y	Y	Y		Y																										
FJ537713	G	G	G	A	G	A	G	C	G	T	C	C	C	C	G	G	A	C	G	A	G	C	G	A	T	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537714	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	G	A	C	G	A	G	C	G	A	T	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537715	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	A	G	C	G	C	C	C	T	C	T	A	A	G	T	G	A	C	T	G	C	C	G	T	C	C	G	T	T	A	T	A	T	C
FJ537716	G	G	G	T	G	A	G	C	G	T	C	C	C	T	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	G	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	A	T	C
FJ537717	G	G	G	T	G	A	G	C	G	T	C	C	C	C	G	G	G	C	G	A	G	C	G	C	C	C	T	C	T	A	A	G	C	A	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537718	G	T	G	T	G	A	T	T	G	T	C	C	C	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537719	G	G	G	A	G	A	G	C	G	C	C	T	C	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	G	A	T	A	C	C
FJ537720	G	G	G	A	G	A	G	C	G	C	C	T	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	G	A	T	A	C	C
FJ537721	G	T	G	A	A	A	G	C	G	T	T	C	C	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	T	C	C	G	A	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	C	C
FJ537722	G	G	G	T	G	A	G	C	G	T	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	C	C
FJ537723	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	G	A	T	A	C	C
FJ537724	G	G	G	T	G	A	G	C	G	C	C	C	C	T	G	G	G	C	G	G	G	C	G	A	T	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	G	C	C	G	T	T	G	C	A	T	C
FJ537725	G	G	G	A	G	A	G	C	G	C	C	T	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537726	G	G	G	T	G	A	T	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	C	C	T	A	T	C	C	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537727	G	G	G	A	G	A	T	C	G	C	C	C	C	T	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	T	C	C	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537728	G	G	G	T	G	A	G	C	G	C	C	C	C	T	G	G	G	C	G	G	G	C	G	A	C	C	C	C	T	A	T	C	C	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537729	G	G	G	T	G	A	G	C	G	T	C	C	C	C	G	G	A	C	G	A	G	C	G	A	T	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537730	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	G	A	C	G	A	G	C	G	A	T	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537731	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537732	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537733	G	G	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	G	T	G	A	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	C	C
FJ537734	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537735	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537736	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537737	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537738	G	T	G	A	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537739	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537740	G	T	G	A	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537741	G	T	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	A	C	T	C	G	C	C	T	T	T	G	C	A	T	C
FJ537742	G	G	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	G	T	G	A	C	T	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537743	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537744	G	T	G	A	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537745	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	G	T	G	A	C	T	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537746	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	G	T	G	A	C	T	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537747	G	G	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	T	G	C	C	G	G	C	C	G	T	T	A	T	A	T	C
FJ537748	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537749	G	G	G	A	G	A	G	C	G	C	C	C	C	C	G	G	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537750	G	G	G	T	G	A	G	C	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537751	A	G	G	T	G	A	G	C	G	T	T	C	C	C	G	G	A	C	A	G	G	C	G	A	T	C	T	T	T	A	A	G	T	G	A	C	T	G	C	C	G	G	C	C	G	T	T	G	C	A	T	C
FJ537752	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537753	G	G	G	T	G	A	G	C	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	C	C	C	A	A	G	T	G	A	A	C	G	C	C	C	G	C	C	T	T	T	A	T	A	C	C
FJ537754	G	G	G	T	G	A	G	C	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	C
FJ537755	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	A	A	C	G	G	G	C	G	A	T	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537756	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537757	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537758	G	G	G	T	G	A	G	C	G	T	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537759	G	G	G	T	G	A	G	C	G	C	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	C	G	A	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537760	G	G	G	T	G	A	G	C	G	T	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	T	T	C	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	C
FJ537761	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	G	A	C	A	A	G	C	G	A	T	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537762	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	G	A	C	A	A	G	C	G	A	T	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537763	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537764	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537765	G	G	G	A	G	A	G	C	G	T	C	C	C	C	G	G	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537766	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537767	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	C
FJ537768	G	T	G	A	G	A	G	C	G	C	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	C	C	C	A	A	G	T	G	A	C	C	G	C	C	C	G	C	C	T	T	T	A	T	A	C	C
FJ537769	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	T	T	C	T	A	A	C	C	G	A	C	C	G	C	C	C	G	T	C	T	T	T	A	T	A	C	T
FJ537770	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	A	C	C	G	C	C	C	G	T	C	T	T	T	A	T	A	C	T
FJ537771	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	T	T	C	T	A	A	C	C	G	A	C	C	G	C	C	C	G	C	C	T	T	T	A	T	A	C	T
FJ537772	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	A	C	C	G	C	C	C	G	T	C	T	T	T	A	T	A	C	T
FJ537773	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	T	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537774	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	T	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537775	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	G	G	A	T	A	T	C
FJ537776	G	G	G	A	G	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	G	A	T	A	C	C
FJ537777	G	G	G	A	G	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	A	C	G	T	C	A	G	T	T	G	T	A	T	C
FJ537778	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537779	G	G	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537780	G	G	G	T	G	C	G	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537781	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	A	C	C	G	A	C	G	T	C	A	G	T	T	G	T	A	T	C
FJ537782	G	T	G	T	G	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	A	C	C	G	A	C	G	T	C	A	G	T	T	G	T	A	T	C
FJ537783	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537784	G	G	G	T	G	A	G	C	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537785	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537786	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537787	G	G	G	A	G	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	G	A	T	A	C	C
FJ537788	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	G	A	T	A	C	C
FJ537789	G	G	G	T	G	A	G	C	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	A	C	T	C	G	C	C	T	T	T	G	C	A	T	C
FJ537790	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	G	T	G	A	C	T	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537791	G	G	G	A	G	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	G	A	T	A	C	C
FJ537792	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	G	G	A	T	A	T	C
FJ537793	G	T	G	T	G	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	G	C	G	T	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537794	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537795	G	G	G	A	G	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	T	T	C	T	A	A	C	C	G	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537796	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537797	G	G	G	A	A	A	G	T	A	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	G	G	A	T	A	C	C
FJ537798	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537799	G	G	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	T	T	C	T	A	A	C	C	G	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537800	G	G	G	A	G	A	G	T	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	T	T	C	T	A	A	C	C	G	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537801	G	G	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	C	C	T	G	A	C	C	G	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537802	G	T	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	C	C	T	G	A	C	C	G	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537803	G	G	A	T	G	A	G	C	A	C	C	C	C	C	G	G	A	C	A	G	G	C	G	A	T	C	T	T	T	A	A	C	C	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537804	G	G	G	T	G	A	G	C	G	T	T	C	C	C	G	G	A	C	A	G	G	C	G	A	T	C	T	T	T	A	A	C	C	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537805	A	T	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	T	A	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537806	G	G	A	T	G	A	G	C	A	C	C	C	C	C	G	G	G	C	G	G	G	T	A	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537807	G	G	A	T	G	A	G	C	A	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537808	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537809	G	T	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	G	T	G	A	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	C	T
FJ537810	G	T	G	T	G	A	T	T	G	T	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	C	T
FJ537811	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537812	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537813	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537814	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537815	G	G	G	T	G	A	G	C	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537816	G	T	G	T	G	A	T	T	G	T	T	C	C	C	G	G	A	C	G	G	G	C	G	A	C	C	C	C	T	G	A	G	C	A	A	C	C	A	C	T	C	G	C	C	T	T	T	G	C	A	T	C
FJ537817	G	G	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	T	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537818	G	G	G	A	G	A	T	T	G	T	C	C	C	C	G	G	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537819	G	G	G	A	G	A	G	T	A	C	C	C	T	C	G	G	A	T	G	G	G	C	G	A	C	C	C	C	T	G	A	C	C	G	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537820	G	G	G	T	G	A	G	C	A	C	C	C	T	C	G	G	A	T	G	G	G	C	G	A	C	C	C	C	T	G	A	C	C	G	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537821	G	G	G	T	G	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	C	C	T	A	A	G	C	A	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	C
FJ537822	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537823	G	G	G	A	A	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	G	A	T	A	C	C
FJ537824	G	G	G	A	G	A	T	T	G	C	C	C	T	C	G	G	A	C	G	A	G	C	G	C	C	C	C	C	T	A	A	C	C	G	A	C	C	G	C	T	C	T	C	C	T	T	T	A	T	A	C	T
FJ537825	G	G	G	A	A	A	G	C	G	C	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	T	C	G	C	C	T	T	G	A	T	A	C	C
FJ537826	G	G	G	A	G	A	T	T	G	C	C	C	T	C	G	G	A	C	G	A	G	C	G	C	C	C	C	C	T	A	A	C	C	G	A	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537827	G	G	G	A	G	A	T	T	G	C	C	C	C	T	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537828	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537829	G	G	G	T	G	A	G	C	A	C	C	C	C	C	A	G	A	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537830	G	G	G	T	G	A	T	C	G	C	C	C	C	C	G	G	A	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537831	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	C	C
FJ537832	G	T	G	A	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537833	G	T	G	T	G	A	T	T	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537834	G	T	G	T	G	A	T	T	G	T	C	C	C	C	G	G	A	C	A	G	G	C	G	A	C	C	T	T	T	A	A	C	C	G	A	C	C	A	C	T	C	G	C	C	G	T	T	G	C	A	T	C
FJ537835	A	T	G	T	G	A	T	T	G	T	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	C	C	G	T	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537836	G	T	G	T	G	A	T	T	G	T	C	C	C	T	G	A	A	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	G	A	T	A	C	C
FJ537837	G	G	G	T	G	A	G	C	G	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	C	C	T	A	A	G	C	G	T	C	C	G	C	T	C	G	C	C	T	T	T	A	T	A	C	T
FJ537838	G	G	G	T	G	A	T	T	G	T	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537839	G	G	G	A	A	A	G	T	A	C	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537840	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	G	A	C	G	A	G	C	G	A	T	C	T	T	T	A	A	G	T	G	A	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537841	G	G	G	T	G	A	G	C	G	T	T	C	C	C	G	G	A	C	A	G	G	C	G	A	T	C	T	T	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537842	G	T	G	T	G	A	T	T	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	T	C	T	T	T	A	A	G	C	G	A	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537843	G	G	G	T	G	A	G	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	T	T	A	A	G	T	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537844	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	A	C	A	T	C
FJ537845	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	C	A	A	C	C	G	A	C	C	G	C	C	G	T	C	C	T	T	T	G	C	A	C	C
FJ537846	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	T	C	T	C	T	A	A	C	C	G	A	C	C	G	C	C	C	G	C	C	T	T	T	A	T	A	C	T
FJ537847	G	T	G	T	G	A	T	T	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	C	C	G	T	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537848	G	T	G	T	G	A	T	T	G	T	C	C	C	C	A	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	C	C	G	T	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537849	G	T	G	T	G	A	T	T	G	T	C	C	C	C	A	G	G	C	G	A	G	C	G	A	C	C	T	T	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	G	T	C
FJ537850	G	T	G	T	G	A	T	T	G	T	C	C	T	C	G	A	A	C	G	G	G	C	G	A	C	C	T	C	T	A	A	C	C	G	T	C	C	G	C	C	C	G	C	C	T	T	T	G	C	A	T	C
FJ537851	G	G	G	T	G	A	G	C	A	C	C	C	C	C	G	G	A	C	G	A	G	C	G	A	T	C	T	T	T	A	A	C	C	G	T	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537852	G	T	G	T	G	A	T	T	G	T	C	C	C	C	G	G	G	C	G	G	G	C	G	A	C	C	T	T	T	A	A	C	C	G	T	C	C	G	C	C	G	T	C	C	T	T	T	A	T	G	C	C
FJ537853	C	G	C	A	G	C	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	C	C	T	A	A	C	C	T	G	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537854	C	G	C	A	G	C	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	C	C	T	A	A	C	C	T	G	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537855	C	G	C	A	G	C	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	A	C	T	A	A	C	C	T	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537856	C	G	C	A	G	C	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	A	C	T	A	A	C	C	T	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537857	C	G	C	A	G	G	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	C	C	T	A	A	C	C	T	G	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537858	C	G	C	A	G	G	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	C	C	T	A	A	C	C	T	G	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537859	C	G	C	A	G	A	G	C	A	G	A	C	C	C	T	C	C	C	A	C	C	G	A	C	T	C	T	C	T	A	A	G	C	T	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537861	C	G	C	A	G	A	G	C	A	G	A	C	C	C	T	C	C	C	A	C	C	G	A	C	T	C	T	C	T	A	A	G	C	T	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537860	C	G	C	A	G	A	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	T	C	T	A	A	G	C	T	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537862	C	G	C	A	G	A	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	T	C	T	A	A	G	C	T	A	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537863	C	G	C	A	G	A	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	T	C	T	A	A	C	C	T	T	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C
FJ537864	C	G	C	A	G	A	G	C	A	G	A	C	C	C	T	T	C	C	A	C	C	G	A	C	T	C	T	C	T	A	A	C	C	T	T	C	C	G	C	C	C	G	C	C	G	T	T	G	C	A	T	C

Table 4.9. Chicken and outgroup genotypes for IFNG at SNP sites polymorphic in the chicken samples.

The synonymous site is listed as "Y". Samples are from Pakistan (FJ537865-FJ537884), Burkina Faso (FJ537885-FJ537904), Senegal (FJ537905-FJ537924), Sri Lanka (FJ537925-FJ537944), Botswana (FJ537945-FJ537964), Bangladesh (FJ537965-FJ537984), Kenya (FJ537985-FJ538004), bamboo partridge (FJ538005-FJ538006), grey francolin (FJ538007-FJ538008), green JF (FJ538009-FJ538010), grey JF (FJ538014, FJ538016), Ceylon JF (FJ538011, FJ538015) and red JF (FJ538012-FJ538013). Bases colours are as per Table 4.7.

Position	1197	1211	1289	1366	1738	1739	1864	1943	2139	2268	2407	2434	3022	3056	3104	3126	3192	3435	4224	4246	4310	4350	4355	4372	4381	4395	4439	4441	4602	4636	4686	4697	4698	4808	4899	4914	4932	4936	4939	4942	5072	5076	5082	5106	5286	5445	6344	6415	6455	6493	6538	6586	6595	6613	6622	6636	6653	6676	6726	6777	6785	6807	6834	6839	6859	6869	7047	7048
Synonymous																																														Y																					
FJ537865	G	T	A	A	C	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	A	C	G	A	G	C	C	C	C	T	C	G	C	G	C	C	C	A	C	C	C	C	G	T	G	C	C	G	T	A	G	C	G	A	T	G
FJ537866	G	T	A	A	C	G	A	C	C	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	A	G	T	G
FJ537867	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	A	C	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537868	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537869	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537870	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537871	G	T	A	G	C	A	A	T	T	G	A	Y	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537872	G	T	A	G	C	A	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537873	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537874	C	T	A	A	C	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	T	C	C	G	C	C	C	C	C	A	G	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537875	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537876	G	T	A	A	C	G	A	T	C	G	A	T	T	G	G	A	A	C	T	C	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	Y	C	G	T	G	C	C	A	C	A	T	G	G	G	C	G
FJ537877	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537878	G	T	A	A	C	A	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	G	T	G
FJ537879	G	T	A	A	C	G	A	T	C	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	A	C	A	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537880	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537881	G	T	A	G	C	G	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	A	C	A	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537882	C	T	A	A	C	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	A	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537883	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	A	C	G	A	G	C	C	A	C	T	C	G	C	G	C	C	C	A	C	C	C	C	G	T	G	C	C	G	T	A	G	C	G	A	T	G
FJ537884	G	T	A	A	C	A	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	A	C	G	A	G	C	C	C	C	T	C	G	C	G	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537885	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537886	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537887	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537888	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537889	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537890	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537891	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537892	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537893	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537894	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537895	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537896	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537897	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537898	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537899	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537900	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537901	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537902	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	Y	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537903	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	A	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537904	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537905	G	T	A	G	C	A	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537906	G	T	A	A	C	A	G	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537907	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537908	G	C	A	A	C	G	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	A	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537909	G	T	A	G	C	A	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537910	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537911	G	T	A	A	C	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	A	C	A	T	A	G	G	A	A	C	A
FJ537912	G	T	A	A	C	G	A	T	C	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	G	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537913	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537914	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	A	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537915	G	T	A	G	C	G	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	A	G	C	A	C	C	C	G	C	G	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537916	G	T	A	A	C	G	A	T	C	G	A	T	T	G	G	A	A	C	C	G	A	C	C	G	T	A	C	T	G	G	A	C	C	T	G	C	A	C	A	C	C	C	C	G	C	G	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	G	C	A
FJ537917	G	T	A	A	C	G	A	T	C	G	A	T	T	G	G	A	A	C	C	G	A	C	C	G	T	A	C	T	G	G	A	C	C	T	G	C	A	C	A	C	C	C	C	G	C	G	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537918	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537919	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	A	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537920	G	T	A	G	C	G	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537921	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537922	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537923	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537924	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537925	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537926	G	T	A	A	C	A	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	A	C	G	A	G	C	C	C	C	T	C	G	C	G	C	C	C	A	C	C	C	C	G	T	G	C	C	G	T	A	G	C	G	A	T	G
FJ537927	G	T	A	A	C	G	A	C	C	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	A	G	T	G
FJ537928	G	T	A	A	C	A	A	T	Y	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537929	G	T	A	G	C	G	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537930	G	C	A	A	C	G	A	C	T	G	A	T	C	G	G	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537931	G	T	A	A	C	G	A	T	C	G	A	T	T	G	G	A	A	C	T	C	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	G	C	A
FJ537932	G	T	A	A	C	A	A	T	Y	G	A	T	Y	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537933	G	T	A	A	C	C	A	T	T	G	G	C	C	C	G	A	A	C	C	G	C	C	G	G	T	A	A	C	G	G	G	T	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537934	G	T	A	A	C	G	A	T	C	G	A	T	T	G	G	A	A	C	C	G	A	C	C	G	T	A	C	T	A	G	G	C	A	C	G	A	G	C	C	C	C	C	C	G	C	G	C	C	C	A	C	C	C	C	G	T	G	C	C	G	T	A	G	G	G	A	C	A
FJ537935	G	T	A	G	C	G	A	T	C	G	A	T	T	G	G	A	A	A	C	G	C	C	G	G	T	A	A	C	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	A	C	C	C	C	G	T	G	C	C	G	T	A	G	C	G	A	T	G
FJ537936	G	T	A	A	C	G	A	T	T	G	G	C	C	C	G	A	A	C	C	G	C	C	G	G	T	A	A	C	G	G	G	T	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537937	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	A	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537938	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537939	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537940	G	T	A	A	C	G	A	T	C	G	A	T	T	G	G	A	A	C	T	C	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	G	C	A
FJ537941	G	T	A	A	C	G	A	C	C	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537942	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537943	G	T	A	A	C	G	A	C	C	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	A	G	T	G
FJ537944	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537945	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537946	G	T	A	A	C	A	A	T	T	G	A	T	C	C	G	A	A	C	C	G	C	C	G	G	T	A	A	C	G	G	G	T	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	A	C	A	C	C	C	C	G	T	G	C	T	G	T	A	G	C	G	A	T	G
FJ537947	G	T	A	A	C	G	A	T	C	G	A	T	T	G	G	A	A	C	T	C	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	G	C	A
FJ537948	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537949	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537950	G	C	A	A	C	G	A	C	T	G	A	T	C	G	G	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	A	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537951	G	T	A	A	C	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	G	C	A	C	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	A	C	A	T	A	G	G	A	A	C	A
FJ537952	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537953	G	T	A	G	C	G	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	A	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537954	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537955	G	T	A	A	C	G	A	T	C	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537956	G	T	A	A	C	A	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537957	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537958	G	C	A	A	C	G	A	C	T	G	A	T	C	G	G	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537959	G	T	A	A	C	G	A	T	C	G	A	T	T	G	G	A	A	C	T	C	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	G	C	G
FJ537960	G	T	A	A	C	G	A	T	C	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537961	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537962	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	C	G	G	T	A	A	C	G	G	G	T	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	A	C	A	C	C	C	C	G	T	G	C	T	G	T	A	G	C	G	A	T	G
FJ537963	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537964	G	C	A	A	C	G	A	C	T	G	A	T	C	G	G	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537965	G	T	A	A	C	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	G	A	T	C	T	C	T	T	G	C	C	G	C	G	T	G	A	A	C	A
FJ537966	C	T	A	A	C	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	T	C	C	G	A	G	G	C	A	C	C	T	A	C	C	C	C	C	A	T	C	T	C	G	T	G	A	C	A	C	G	T	G	A	G	C	A
FJ537967	G	T	G	A	C	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	T	C	C	G	A	G	G	C	A	C	C	T	A	C	C	C	C	C	A	T	C	T	C	G	T	G	A	C	A	C	G	T	G	A	G	C	A
FJ537968	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	T	C	G	C	T	T	T	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537969	G	T	A	A	C	G	A	T	C	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	A	G	T	G
FJ537970	G	T	A	A	C	A	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	A	C	G	A	G	C	C	C	C	T	C	G	C	G	C	C	C	A	C	C	C	C	G	T	G	C	C	G	T	A	G	C	G	A	T	G
FJ537971	G	T	A	A	C	G	A	T	T	G	G	C	C	C	G	A	A	C	C	G	C	C	G	G	T	A	A	C	G	G	G	T	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537972	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537973	G	T	A	A	C	G	A	C	C	G	A	T	C	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	G	A	T	C	T	C	T	A	G	C	C	G	C	G	T	G	A	A	C	A
FJ537974	G	T	A	A	C	G	A	C	C	G	A	T	C	G	A	A	A	C	C	G	C	C	G	G	T	A	A	C	G	G	G	T	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537975	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537976	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537977	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537978	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537979	G	T	A	A	C	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	T	C	C	C	G	A	T	C	T	C	T	A	G	C	C	A	C	A	T	G	G	A	C	A
FJ537980	G	T	A	A	C	G	A	T	T	A	A	T	T	G	G	A	G	C	C	G	C	C	C	G	T	A	A	T	G	G	G	C	A	C	G	A	G	C	A	C	C	C	C	G	T	C	C	C	C	A	T	C	C	C	T	A	G	C	C	G	C	G	T	G	A	A	C	A
FJ537981	G	T	A	A	C	G	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	C	C	G	T	A	A	T	G	G	G	C	A	C	G	A	G	C	C	C	C	C	C	G	T	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537982	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537983	G	T	A	A	C	G	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	C	C	G	T	A	A	T	G	G	G	C	A	C	G	A	G	C	C	C	C	C	C	G	T	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537984	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537985	G	T	A	A	C	G	A	T	C	G	A	T	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	T	C	G	C	G	C	C	C	A	C	C	C	C	G	T	G	C	C	G	T	A	G	C	G	A	T	G
FJ537986	G	T	A	A	C	A	A	T	T	G	A	T	C	G	A	T	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	A	C	G	A	G	C	C	C	C	T	C	G	C	G	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537987	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	A	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537988	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537989	G	T	A	A	C	G	A	C	C	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	G	C	A
FJ537990	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537991	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	A	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537992	G	T	A	A	C	G	A	Y	C	G	A	Y	C	G	A	T	A	C	C	G	C	C	C	A	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	T	T	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537993	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537994	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537995	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537996	G	T	A	G	T	G	A	T	T	G	A	T	C	G	A	A	A	C	C	G	C	C	C	G	T	A	A	C	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ537997	G	T	A	A	C	G	A	C	C	G	A	T	C	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	G	A	T	C	T	C	T	A	G	C	C	G	C	G	T	G	A	A	C	A
FJ537998	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	A	C	C	G	T	A	C	T	G	G	A	C	C	T	G	C	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ537999	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538000	G	C	A	A	C	G	A	C	T	G	A	T	C	G	G	A	A	C	C	G	C	T	C	G	T	C	C	T	G	G	G	C	C	C	G	A	G	G	C	A	C	C	C	G	C	C	C	T	C	G	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ538001	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538002	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538003	G	T	A	A	C	G	A	T	C	G	A	T	T	G	G	A	A	C	C	G	A	C	C	G	T	A	C	T	A	G	G	C	A	C	G	A	G	C	C	C	C	C	C	G	C	G	C	C	C	A	C	C	C	C	G	T	G	C	C	G	T	A	G	G	G	A	C	A
FJ538004	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	A	C	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A
FJ538005	G	T	A	A	C	A	A	T	G	G	A	T	T	G	G	A	A	A	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	A	C	C	C	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538006	G	T	A	A	C	A	A	T	G	G	A	T	T	G	G	A	A	A	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	A	C	C	C	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538007	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	G	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	A	C	C	C	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538008	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	A	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	A	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538009	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	G	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	G	C	C	T	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538010	G	T	A	A	C	A	A	T	T	G	A	T	T	G	G	A	A	G	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	G	C	C	T	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538014	G	T	A	A	C	G	A	C	T	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	C	T	G	G	G	C	C	C	G	A	G	C	C	C	C	C	C	G	C	G	C	C	C	G	T	C	C	C	G	T	G	C	C	A	C	A	T	G	A	A	C	A
FJ538016	G	T	A	A	C	G	A	C	T	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	C	T	G	G	G	C	C	C	G	A	G	C	C	C	C	C	C	G	C	G	C	C	C	G	T	C	C	C	G	T	G	C	C	A	C	A	T	G	A	A	C	A
FJ538011	G	T	A	A	C	G	A	C	T	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	C	T	G	G	G	C	C	C	G	A	G	C	A	C	C	C	C	G	C	G	C	C	C	G	T	C	C	C	G	T	G	C	C	G	C	A	T	G	A	G	C	A
FJ538015	G	T	A	A	C	G	A	C	T	G	A	T	C	G	G	A	A	C	C	G	C	C	C	G	T	A	C	T	G	G	G	C	C	C	G	A	G	C	A	C	C	C	C	G	C	G	C	C	C	G	T	C	C	C	G	T	G	C	C	G	C	A	T	G	A	G	C	A
FJ538012	G	T	A	A	C	A	A	T	T	A	A	T	C	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	C	C	C	A	C	A	T	G	G	A	C	A
FJ538013	G	T	A	A	C	G	A	T	T	G	A	T	C	G	G	A	A	C	C	G	C	T	C	G	G	A	C	T	G	G	G	C	C	C	A	A	A	C	A	C	C	C	C	G	C	C	C	C	C	G	T	C	C	C	G	T	G	C	C	A	C	A	T	G	G	A	C	A

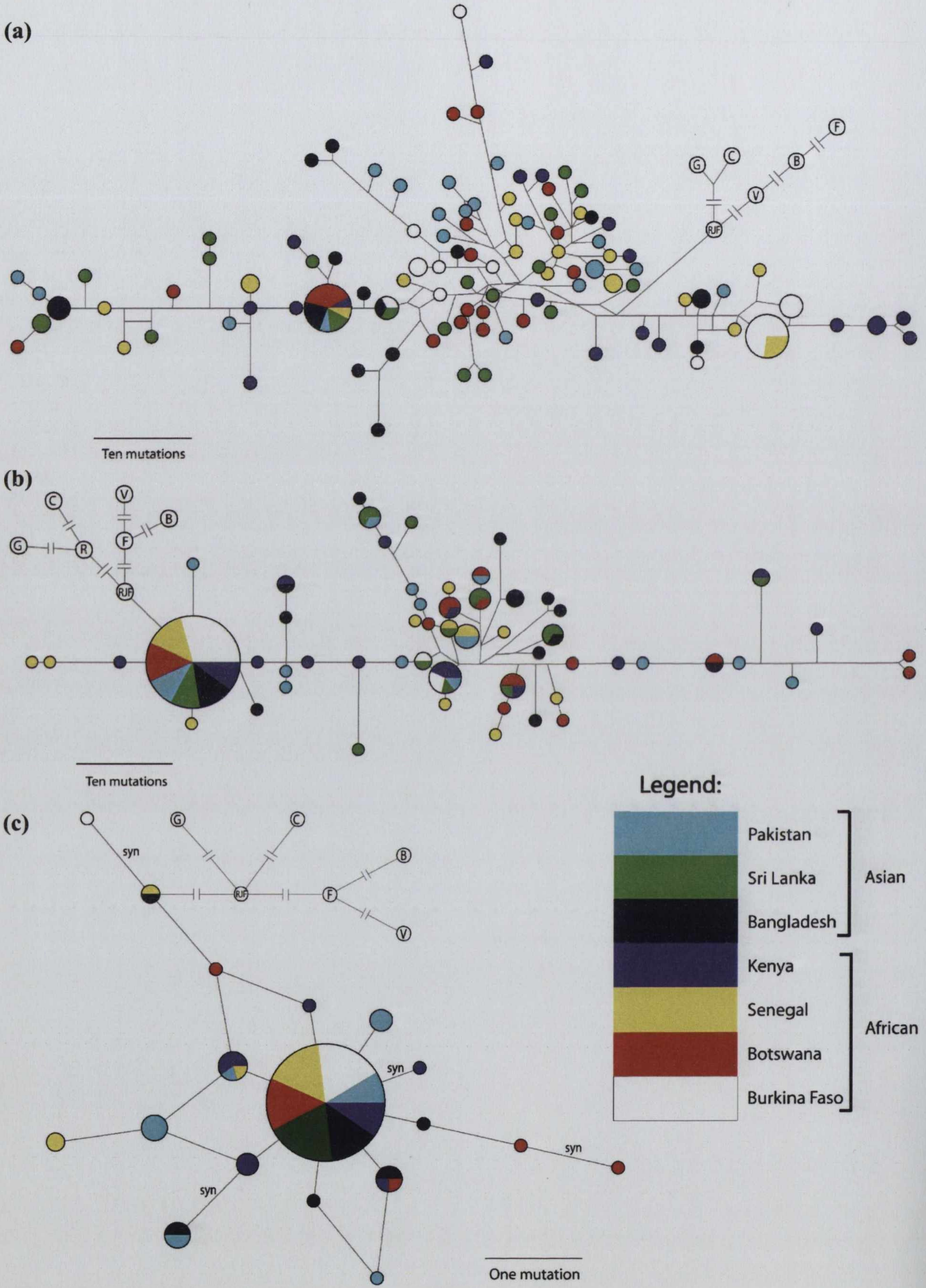
Further evidence for the high diversity at IL1B was observed in the number of haplotypes illustrated in a median-joining network (Figure 4.4a). In contrast, the IFNG network (Figure 4.4b) was considerably less diverse, and had one numerically dominant haplotype. Networks of both genes showed little observable association of haplotypes with geography, though IL1B did have one specific African branch with a major Burkina Faso-Senegal haplotype present.

When a network was constructed for IL1B using coding SNPs only (Figure 4.4c), the most significant difference was the single dominant haplotype present among the coding region network. Notably, this was five sequence differences in phylogenetic distance to the red JF genome sequence.

In each network the red JF genome sequence was the most closely related species to the domestic chicken samples compared to the other JF. At IFNG, the red JF was one synonymous SNP in phylogenetic distance to the largest haplotype in the chicken samples, implying that red JF was the most likely source of diversity at IFNG.

The different functional predictions made by SIFT, Polyphen and PMut for the nonsynonymous SNPs between red JF and the chicken populations at IL1B showed three nonsynonymous SNPs segregating at high frequency (Table 4.10). In a T-Coffee alignment (not shown) (Notredame et al. 2000) of all available IL1B protein sequences, all the red JF amino acid variants at polymorphic sites in chicken also have occurred in the turkey, duck, goose and pigeon (with the sole exception of S164 for the latter), which suggested the chicken alleles were derived.

Figure 4.4. Median-joining networks of chicken population haplotypes for (a) IL1B, (b) IFNG and (c) IL1B for coding SNPs only.



Legend to Figure 4.4: Branch lengths are proportional to the number of sequence differences between haplotypes. The outgroup samples are represented by the colourless nodes. V represents the green JF sequences; F the grey francolin; B the bamboo partridge; G the grey JF; C the Ceylon JF; R the red JF; and RJF the genome sequence. For IL1B (a) and (c) there were no differences between the genome sequence and the red JF sample. In (c) synonymous SNPs between chicken samples are denoted as "syn"; the rest were nonsynonymous.

Table 4.10. Predicted functional impacts of nonsynonymous SNPs among chickens on the IL1B protein product by SIFT, Polyphen and PMut.

Gene		Amino Acid						
Position	P [1]	Red JF	DA [2]	SIFT	Polyphen	PMut	N [3]	Outcome
225	22	G	S	tolerated	possibly damaging	n/d [4]	3	n/d
630	103	Y	F	tolerated	benign	neutral	135	neutral
630	103	Y	S	tolerated	probably damaging	n/d [4]	5	n/d
633	104	A	V	tolerated	benign	neutral	2	neutral
939	150	R	G	deleterious	benign	n/d [4]	5	n/d
952	154	A	E	deleterious	benign	n/d [4]	127	n/d
952	154	A	G	deleterious	benign	neutral	13	prob. neutral
960	157	R	G	deleterious	benign	n/d [4]	2	n/d
982	164	S	I	deleterious	benign	neutral	137	prob. neutral
982	164	S	T	tolerated	benign	neutral	3	neutral
1063	191	E	G	deleterious	benign	n/d [4]	2	n/d
1261	230	S	F	deleterious	probably damaging	pathological	15	deleterious

[1] Amino acid site. [2] Derived allele – in some cases this was the most frequent in the chicken samples, and at positions 939, 960 and 1063 was observed in the outgroup samples as well. [3] Number of observed samples with minor allele. [4] Not determined.

4.3.2 Summary statistics and tests of neutrality:

The summary statistics and tests of neutrality (Table 4.11) illustrated further contrasts in diversity at the two genes. IL1B had 105 haplotypes among 140 genotypes, so its haplotype diversity (*Hd*, Equation 2) was significantly high, while IFNG had 56 haplotypes in 140 genotypes and thus its *Hd* was significantly low. The nucleotide diversity (π) at IL1B was much higher than at IFNG. Interestingly, the relative numbers of haplotypes in each continent were much lower than for both continents together. This indicated a high number of unique haplotypes, highlighting the high population-level diversity. This was reflected in the Fu's F_S (Equation 11) values that showed an excess of rare alleles as a result of the high number of unique haplotypes when both continents were examined in tandem. Nucleotide diversity tended to be higher in Asia (3.58 per kb) than in Africa (2.85 per kb) at IFNG, but the relative difference was reduced at IL1B.

Table 4.11. Gene data, summary statistics and tests of neutrality from DnaSP for IL1B and IFNG.

Test	SNPs	H [2]	*Hd* [3]	π [4]	θ_w [5]	Tajima's *D*	Fu & Li's		Fay & Wu's *H*	Fu's F_s
							D	*F*		
All IL1B [1]	52	104	0.99	4.92	4.01	0.83	1.63	1.50	-7.40	-98.55
P value			*0.018*	*n/s*	*0.040*	*0.020*	*<0.001*	*<0.001*	*0.011*	*0.015*
Asia IL1B	44	51	0.99	4.97	4.01	0.80	1.82	1.71	-4.21	-28.70
P value			*n/s*	*n/s*	*n/s*	*0.021*	*0.002*	*0.004*	*n/s*	*n/s*
Africa IL1B	44	56	0.97	4.76	3.78	1.04	0.84	1.03	-6.941	-32.49
P value			*n/s*	*n/s*	*n/s*	*0.006*	*n/s*	*n/s*	*0.011*	*n/s*
All IFNG [1]	68	56	0.82 *	3.18	3.03	0.15	1.06	0.76	2.47	-14.23
P value			*<0.001*	*n/s*	*n/s*	*n/s*	*n/s*	*n/s*	*n/s*	*0.048*
Asia IFNG	61	33	0.85 *	3.58	3.11	0.38	0.85	0.77	4.20	-4.71
P value			*<0.001*	*n/s*	*n/s*	*n/s*	*n/s*	*n/s*	*n/s*	*0.007*
Africa IFNG	61	33	0.76 *	2.85	2.78	0.08	0.62	0.48	4.20	-4.11
P value			*<0.001*	*n/s*	*n/s*	*n/s*	*n/s*	*n/s*	*n/s*	*0.019*

The length sequenced at each locus was 2,351 bp for IL1B and 4,066 bp for IFNG. [1] All samples are constituted by the Asian and African samples. [2] Number of haplotypes. [3] Haplotype diversity. [4] Mean number of pairwise differences per kb between sequences. [5] Watterson's estimator per kb. P values are generated by coalescent simulations for given recombination using DnaSP; only those with $p < 0.05$ are given. P values are significantly high except where stated. * Value significantly low.

There was a strong and consistent contrast in the outcome of the tests of selection at each gene. The Tajima's *D* value (Equation 7) for IL1B was significantly high because of the excess of intermediate relative to low frequency alleles present; IFNG had a neutral value. Likewise, Fu and Li's *D* (Equation 8) and *F* (Equation 9) statistics were extreme at IL1B but neutral at IFNG – notably, the deficit of singletons was stronger in Asia than Africa at IL1B. Fay and Wu's *H* at IL1B suggests that there was a significant excess of high-frequency derived alleles compared to intermediate ones; for IFNG it did not. Further tests of Fay and Wu's *H* (Equation 10) at the coding segment of IL1B reveal outlying values: -6.49 ($p < 0.001$) for all, -4.40 ($p < 0.001$) for Asia and -4.62 ($p < 0.001$) for Africa. This suggested that the relative excess of high-frequency alleles encompasses the protein-coding portion of IL1B.

4.3.3 Neutrality of pairwise difference distributions:

The distribution of the numbers of pairwise differences (π) between sequences can yield information about how a gene's diversity has changed in the past in terms of both demography and selective effects (Rogers & Harpending 1992). Historical population sizes were simulated using DnaSP to fit observed pairwise frequency (π)

data for population size changes (ΔN), where τ is the timing of ΔN: as τ increases, it becomes more ancient (Figure 4.5). Estimates of the historical timings of the major demographic events were calculated according to Equation 1 (Table 4.12). Notably, the IL1B coding sequence had a much more brief estimated history, suggestive of purifying selection, and was consistent with the possibility that selective sweeps have reduced variability in this region as well.

Table 4.12. Estimated time of the peak gene population expansions (kya).

Gene	Estimated τ	Time (t)
IL1B	11.2	1,323
	15	1,772
IFNG	6.22	425
	17	1,161
IL1B CDS	0.174	72.3

τ is the mean peak pairwise difference (π) values simulated in Figure 4.6. t is time in generations where one generation is equal to one year.

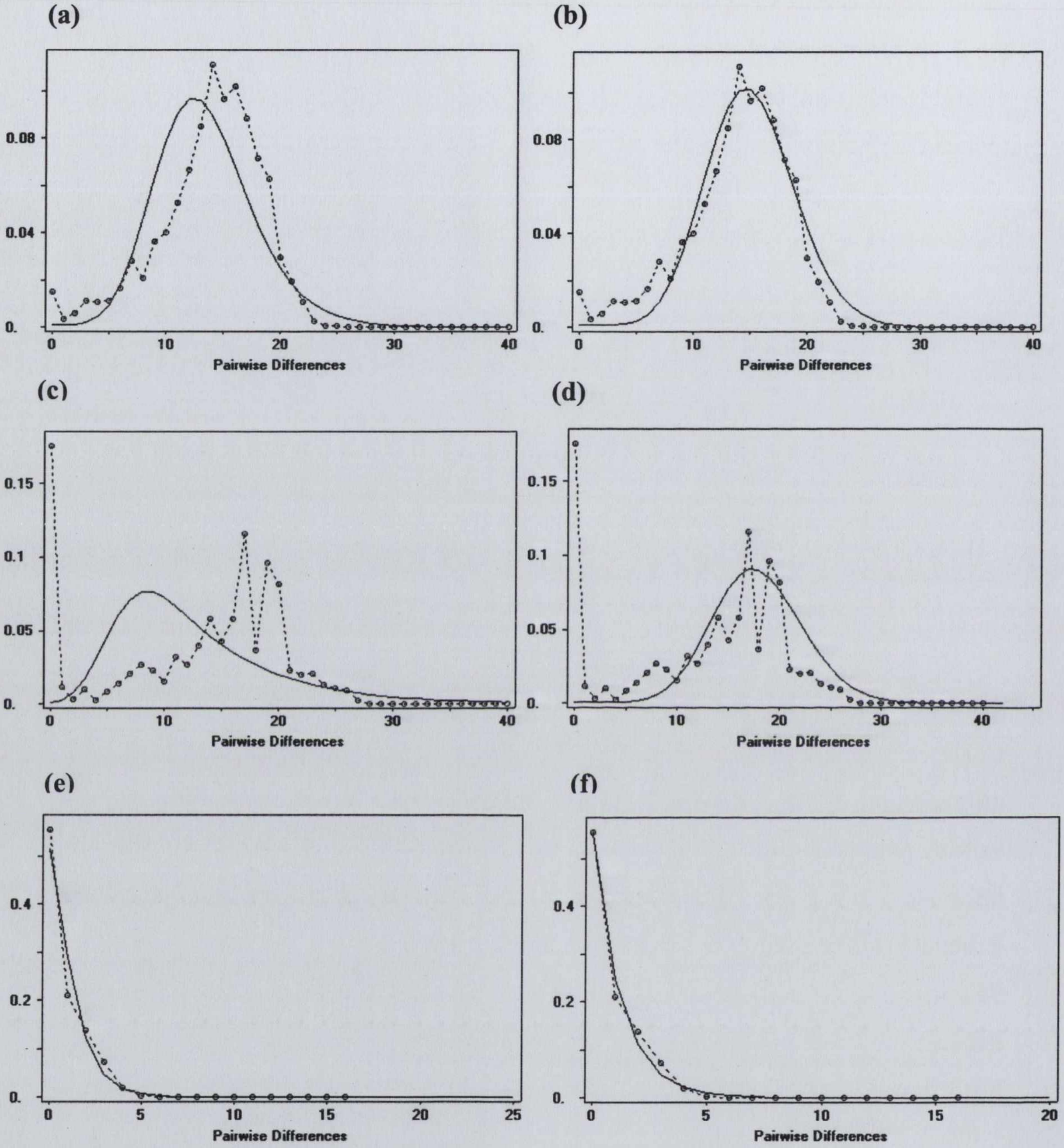
4.3.4 Coalescent simulations:

Coalescent simulations examined how different models of demographic history and migration might affect the distribution of observed diversity at a locus. Despite varying parameters, including values consistent with domestication and resequencing information, the genotyped values of π and Tajima's D at IL1B consistently lie at the extreme positive end of the simulated scanMS distribution: all values lie in 96.9 to 99.9 percentiles for all models (Table 4.13). In contrast, the observed IFNG data was extreme only with migration and where population size was constant during part of its history.

4.3.5 Recombination rates:

There was a clear disparity in the recombination rates between the genes (Table 4.14). The recombination rate (R, Equation 6) at IL1B was exceptionally high, and there was a clear relative disparity between R and the minimum number of recombination events (R_M, Equation 5) among IL1B and IFNG. GC content, an indicator of the extent of recombination, was high at IL1B (Duret & Arndt 2008), and Kelly's Z_{nS} (Equation 3) indicated that LD was more disturbed at IL1B.

Figure 4.5. The frequency of pairwise differences in (a,b) IL1B, (c,d) IFNG and (e,f) the IL1B coding region.



(a), (c) and (e) were computer simulated according to the mean pairwise difference value; (b), (d) and (f) were manually adjusted for the peak pairwise difference value. The solid line is the simulated pairwise differences according to the parameters and the dashed is the observed one. The y-axis indicates the frequency of the pairwise difference. For IL1B, (a) $\Delta N = 388$ and $\tau = 11.2$; (b) $\Delta N = 10{,}000$ and $\tau = 15.0$. For IFNG, (c) $\Delta N = 149$ and $\tau = 6.2$; (d) $\Delta N = 1{,}000$ and $\tau = 17.0$ (as per Table 4.12). For the IL1β coding sequence (e) had $\Delta N = 1572$ and $\tau = 0.17$; (f) $\Delta N = 0$ and $\tau = 0$. The length of the sequenced region (m_T) was 2,351 bp for IL1B, 4,066 bp for IFNG and 668 bp for the IL1B coding sequence.

Table 4.13. The results of the coalescent simulations with MS and scanMS for IL1B and IFNG for specific models compared with the observed data.

IL1B	Model	Increasing [1]		Constant [2]		Exponential [3]	
	Migration	Off	On	Off	On	Off	On
Pairwise Differences (Obs = 13.78) [5]	Mean	10.9	11.0	10.8	6.6	10.8	8.5
	Percentile 0.05	8.5	8.8	8.6	4.9	8.2	6.6
	Percentile 0.95	13.4	13.2	13.2	8.5	13.6	10.6
	Std dev [4]	1.5	1.3	1.4	1.1	1.6	1.2
	Obs percentile	97.2	98.3	98.3	99.9	96.9	99.9
Tajima's D (Obs = 0.83) [5]	Mean	0.00	0.02	-0.01	-1.21	-0.03	-0.67
	Percentile 0.05	-0.69	-0.62	-0.65	-1.70	-0.77	-1.22
	Percentile 0.95	0.69	0.66	0.66	-0.68	0.74	-0.10
	Std dev	0.42	0.38	0.40	0.31	0.46	0.34
	Obs percentile	97.9	98.6	98.4	99.9	97.2	99.9

IFNG	Model	Increasing		Constant		Exponential	
	Migration	Off	On	Off	On	Off	On
Pairwise Differences (Obs = 12.92) [5]	Mean	12.3	12.4	12.3	7.5	12.2	9.7
	Percentile 0.05	9.6	10.1	9.7	5.6	8.9	7.6
	Percentile 0.95	15.2	14.8	15.0	9.6	15.5	11.9
	Std dev	1.7	1.4	1.6	1.2	2.0	1.3
	Obs percentile	64.2	64.5	65.1	99.9	63.9	99.4
Tajima's D (Obs = 0.15) [5]	Mean	0.00	0.02	0.00	-1.23	-0.03	-0.68
	Percentile 0.05	-0.68	-0.58	-0.68	-1.71	-0.84	-1.22
	Percentile 0.95	0.73	0.63	0.67	-0.70	0.81	-0.10
	Std dev	0.43	0.37	0.41	0.31	0.50	0.34
	Obs percentile	63.7	63.7	64.2	99.9	63.9	99.3

[1] Model 1 – a constant population size gradual increase (from 10^5 to 10^6). [2] Model 2 – a large constant population size (10^6). [3] Model 3 – no population size change until a recent exponential increase (from 10^5 to 10^6). [4] Standard deviation of the simulated data. [5] Observed values for each statistic. The observed mean pairwise differences (π) and Tajima's D are generated by DnaSP. The observed percentile is the point where the observed value would lie on the scanMS simulated distribution.

Table 4.14. Recombination at IL1B and IFNG according the percentage GC content, Hudson's R and R_M, Kelly's Z_nS per kb and the number of gene conversion tracts (N) from DnaSP, and significant estimated range of ρ from LDhat.

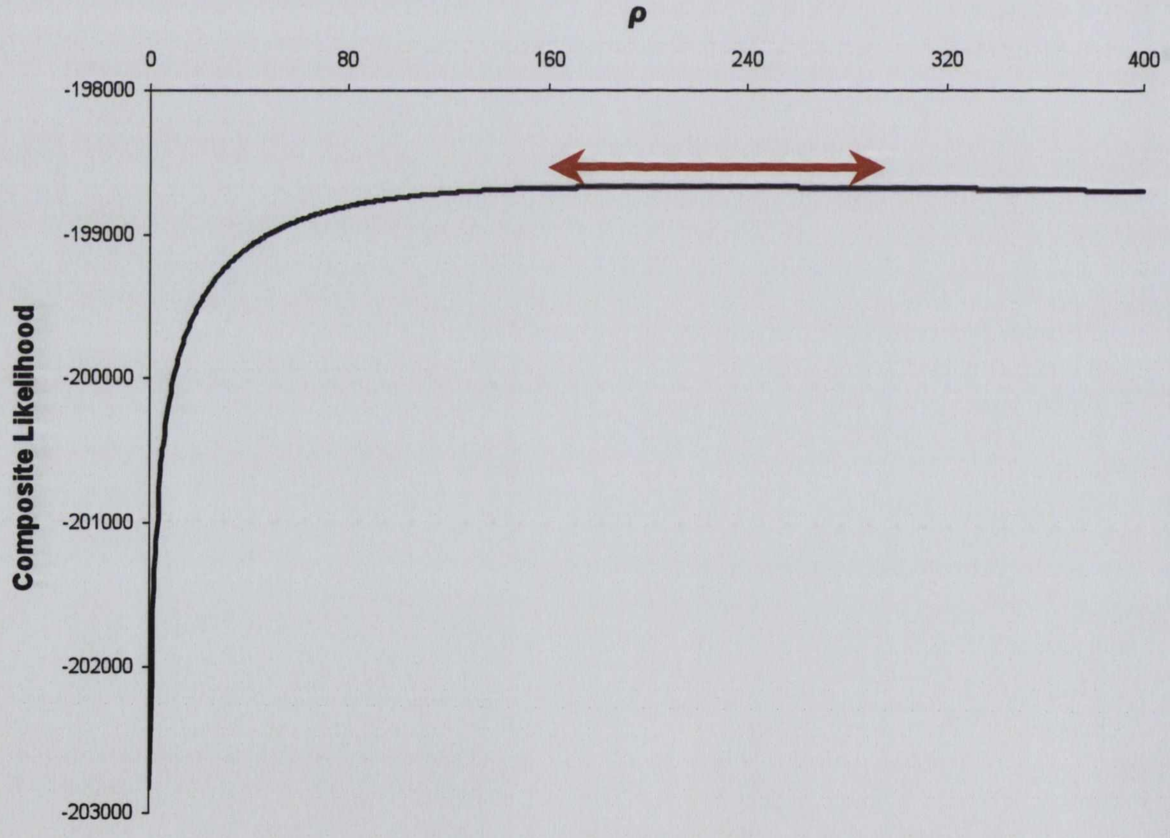
Gene	GC content			R	R_M [1]	Z_{nS} [2]	LDhat ρ	N [3]
	Total	Coding	Non-coding					
IL1B	0.649	0.657	0.646	125.98	23	32.37	158.6 - 293.3	79
IFNG	0.431	0.416	0.433	11.10	17	64.35	41.0 - 55.9	31

[1] For coalescent simulations using DnaSP p = 0.029 for IL1B and p < 0.0001 for IFNG. [2] p = 0.107 for IL1B, p = 0.458 for IFNG. [3] One-tailed χ^2 p < 0.01. [3] Total number of gene conversion tracts identified between all seven populations.

The estimated range of recombination rates by LDhat, although wide, indicate that recombination at IL1B was high and was about three to seven times higher than that at IFNG (Figure 4.6). LDhat was used to examine the local recombination rate at SNP sites: IL1B had a mean SNP-specific ρ (0.594) greater than that of IFNG (0.298), and local recombination was largely uniform across the genes. The LDhat SNP hotspot distribution showed that IL1B's mean heat (0.018) was greater than IFNG's (0.008) and that the hotspot intensities were largely level throughout the genes. Thus higher recombination at IL1B was due to gene-wide effects, rather than local or hotspots phenomena at either gene.

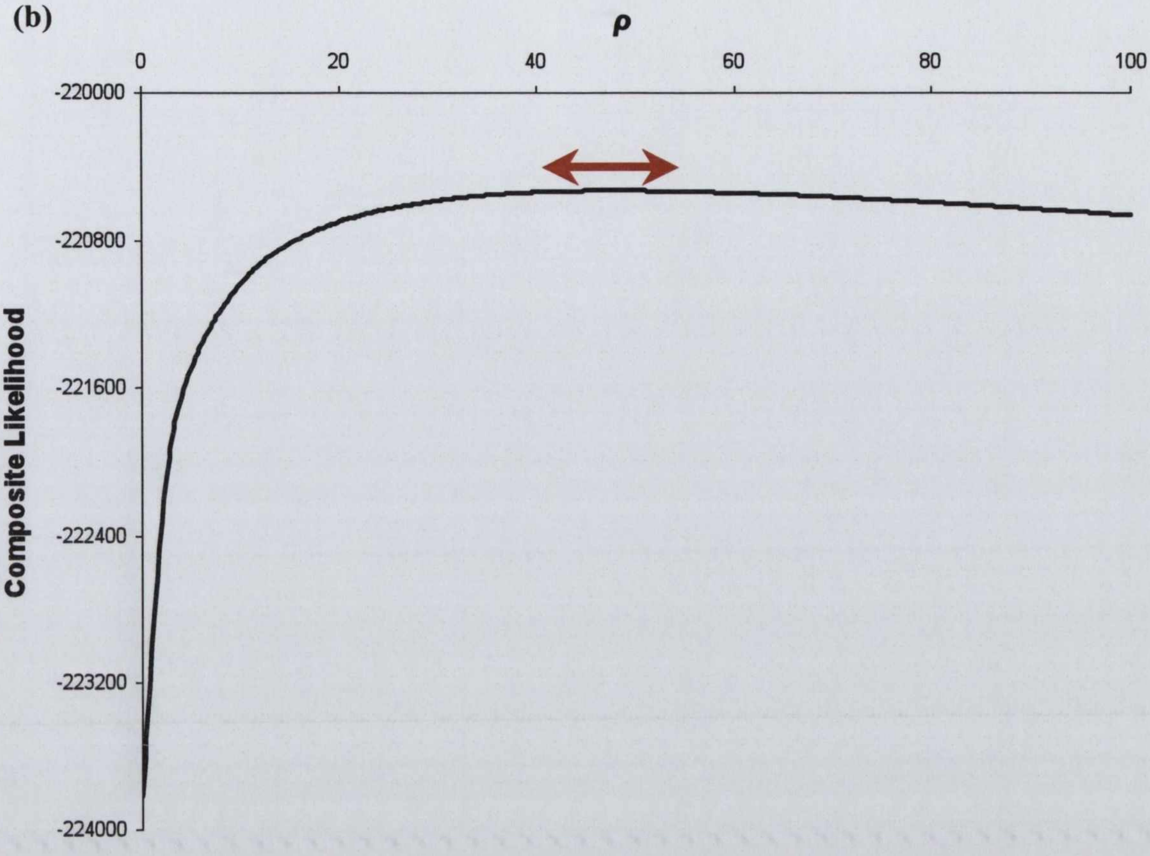
Figure 4.6. Composite maximum likelihood estimators for ρ for (a) IL1B and (b) IFNG.

(a)



The midpoint of the estimate for ρ as determined by the LDhat CLE was 209.5 for IL1B with a significantly more likelihood range of ρ = *158.6* to ρ = *293.3* (χ^2 p < 0.01) as indicated by the red arrow. Note that the x-axis scale for the ρ estimate is longer for IL1B than IFNG.

Figure 4.6. (continued).

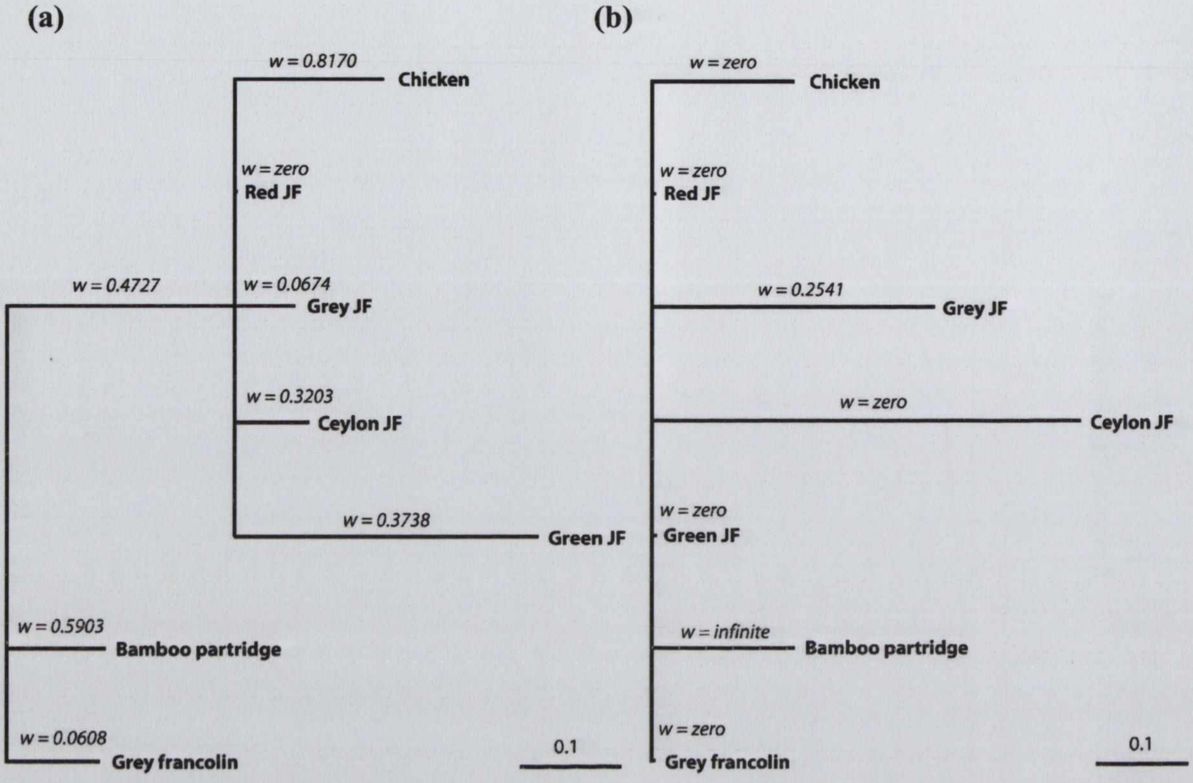


The midpoint of the estimate for ρ as determined by the LDhat CLE was 47.5 for IFNG with a significantly more likely range of $\rho = 41.0$ to $\rho = 55.9$ (χ^2 $p < 0.01$) as indicated by the red arrow. Note that the x-axis scale for the ρ estimate is shorter for IFNG than IL1B.

4.3.6 Interspecies tests for selection:

In order to test specific lineages and sites for selection, tests using PAML were implemented. Using branch lengths and ω values from the free-ratio model, neighbour-joining trees were constructed from coding sequences for each of the seven species sequenced (Figure 4.7). PAML analysis was not informative for IFNG as a consequence of a lack of CDS mutations between species: only seven substitutions among the seven samples are observed at this gene (Table 4.15). For IL1B, the chicken ω value (0.82) was elevated compared to those of the other species (Figure 4.8). Lineage-specific (M2) LRTs examined the likelihood of a model with specific branches under a different ω ratio to others with a neutral model and these results supported the chicken lineage having a ω value higher compared to the other species (Table 4.16).

Figure 4.7. Neighbour-joining phylogenies of (a) IL1B and (b) IFNG.



Branch lengths are estimated by maximum likelihood under the free-ratio model, which assumes an independent ω-ratio for each branch: these are displayed above each branch. The branch lengths displayed are 0.1 of the total branch lengths for that tree.

Table 4.15. Estimated distribution of nonsynonymous ($N.d_N$) and synonymous ($S.d_S$) SNPs by the codeml free-ratio model for samples at IL1B and IFNG.

Gene	IL1B		IFNG	
Sample	$N.d_N$	$S.d_S$	$N.d_N$	$S.d_S$
Chicken [1]	8.1	2.1	0	1.0
Red JF	0	0	0	0
Grey JF	1.0	3.1	1.0	1.0
Ceylon JF	3.0	2.1	0	3.1
Green JF	13.4	7.5	0	0
Bamboo partridge	5.0	1.8	1.0	0
Grey francolin	1.4	4.9	0	0

[1] 10.8 nonsynonymous and 4.8 synonymous changes were in the branch ancestral to the *Gallus* genus.

Table 4.16. Generated PAML parameters used and output for significant test results involving chicken for IL1B.

Lineage(s)	Model	Parameters	Likelihood	$\omega = d_N/d_S$	$2\Delta ML$	P value
All	M1	ω = estimated independently for all	-1253.410	See Figure 4.7(a)	-	-
Gallus genus and ancestral branch [1]	M2 two ratio	ω = estimated	-1255.941	*Gallus*: ω = 0.3864 B, F: ω = 0.2055	9.143	0.0025
		ω = fixed	-1260.512	1		
Gallus genus and ancestral branch [1]	M2 three ratio	ω = estimated	-1255.323	Chicken: ω = 0.8167, All JF: ω = 0.3326, B, F: ω = 0.2082	10.317	0.0013
		ω = fixed	-1260.481	1		
Gallus genus [1]	M2 four ratio	ω = estimated	-1254.477	Chicken: ω = 0.8166, R, G: ω = 0.0673, C, V: ω = 0.3577, B, F: ω = 0.3124	5.275	0.0216
		ω = fixed	-1257.114	1		
Gallus genus [1]	M2 two ratio	ω = estimated	-1256.353	*Gallus*: ω = 0.3604, B, F: ω = 0.3124	8.128	0.0044
		ω = fixed	-1260.417	1		
Gallus genus [1]	M2 three ratio	ω = estimated	-1255.548	Chicken: ω = 0.8166, JF: ω = 0.2843, B, F: ω = 0.3124	9.677	0.0019
		ω = fixed	-1260.386	1		
All [2]	M2a [2]	$\omega_0 = 0$ (81.0%)	-1229.307	$\omega_2 = 4.3622$ (6.60%), $\omega_1 = 1$ (12.34%)	10.739	0.0047
	M1a	$\omega_0 = 0$ (79.3%)	-1234.676	$\omega_1 = 1$ (20.66%)		
All [2]	M8 [3]	$\omega_{0\text{-}7} = 0$ (9.2% each)	-1229.337	$\omega_8 = 0.0867$ (9.23%), $\omega_9 = 0.9999$ (9.23%), $\omega_{10} = 3.9919$ (7.75%)	10.701	0.0047
	M7	$\omega_{0\text{-}7} = 0$ (10.0% each)	-1234.688	$\omega_{8,9} = 1$ (10.00%)		

The M2 models are branch-specific models. The M1a-M2a and M7-M8 comparisons are site-specific models. *2ΔML* is twice the difference between the models' likelihoods. [1] One degree of freedom for LRT. [2] Two degrees of freedom for LRT. [3] BEB analysis (Yang et al. 2005) showed two sites where P(ω > 1) > 97.5%: 48 (ω = 5.428 ± 1.969) and 222 (ω = 5.403 ± 2.002). [4] BEB suggested five sites where P(ω > 1) > 93.8% (see Table 4.17).

Site specific tests between M8 and M7 (and also M2a and M1a) suggest that for IL1B about 93% of sites may have a ω ratio between 0 and 1, but 7% may have a positive ω value (ω = 4.36 for M2a, ω = 3.99 for M8; Table 4.16). In both cases the variable models (M2a, M8) were significantly more likely ($p < 0.01$) than the corresponding neutral models (M1a, M7). BEB analysis for M8 determined five candidates that may be subject to selection (Table 4.17): all these sites have nonsynonymous SNPs segregating between chicken and the outgroup samples. The functional impact at amino acids 51, 75 and 202 was predicted to be neutral (Table 4.18).

Table 4.17. IL1B sites potentially under positive selection according to PAML M8.

Site	Amino Acid	ω	ω SE[1]	$P(\omega>1)$	Bases	Exon	Out[2]	SNPs
48	Arginine	5.413	1.604	0.996	CGG	3	C:	CAG (Glutamine)
							V, B:	CCG (Proline)
							F:	CTG (Leucine)
51	Arginine	5.152	1.917	0.939	CGT	3	V:	GGT (Glycine)
							B, F:	CCG (Serine)
75	Serine	5.148	1.867	0.944	AGC	3	V, B:	TGC (Cysteine)
							F:	CGC (Arginine)
202	Methionine	5.103	1.901	0.936	ATG	6	V, B:	ACG (Threonine)
							F:	AGG (Arginine)
222	Threonine	5.397	1.629	0.992	ACT	6	V, B:	GCT (Alanine)
							F:	CCT (Pro) / GCT (Ala)

[1] Standard error. [2] Outgroup samples: C stands for Ceylon JF, V for green JF, B for bamboo partridge and F for grey francolin.

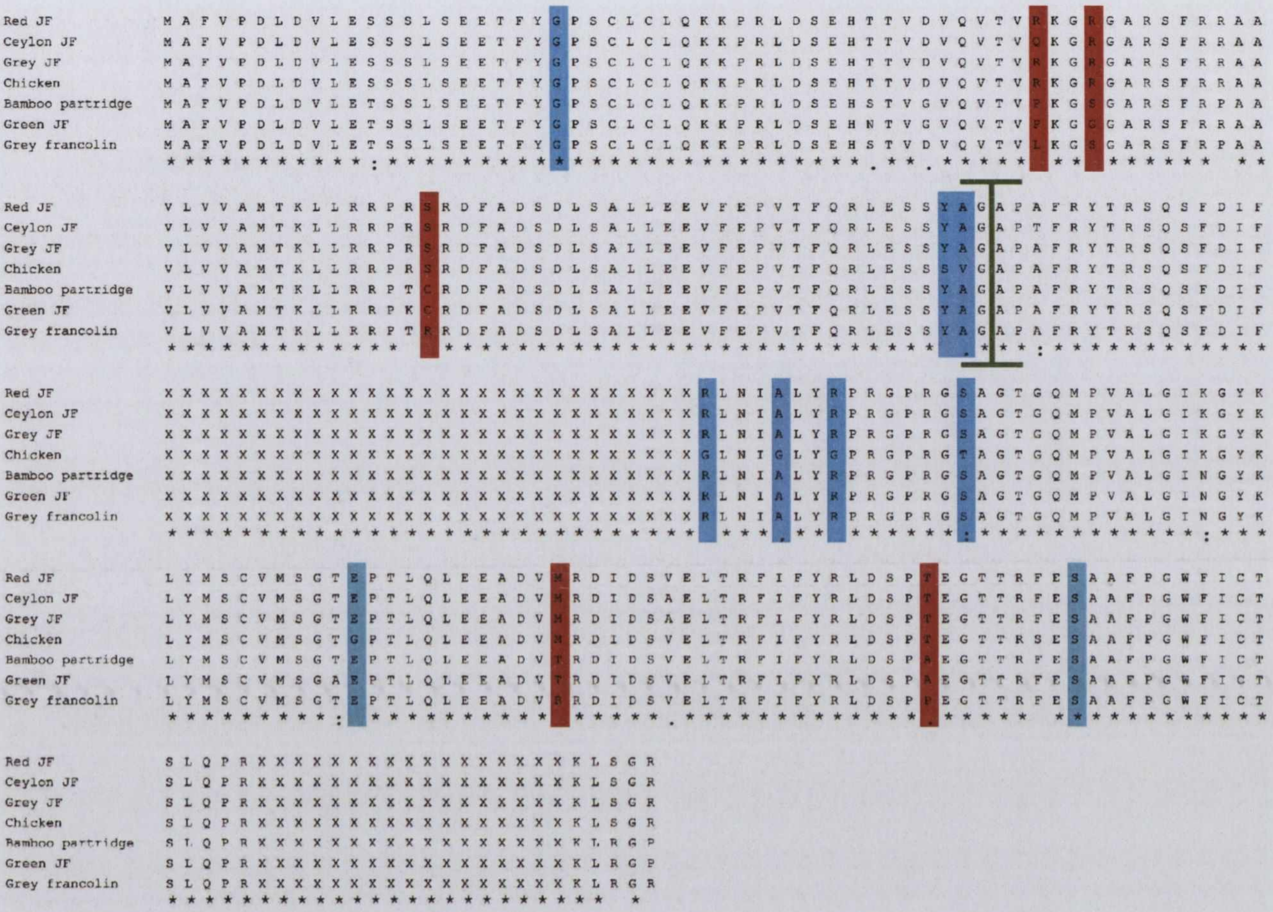
Table 4.18. Predicted functional impacts of outgroup nonsynonymous SNPs for polymorphic candidate sites from M8 BEB on the IL1B protein product.

Gene		Amino Acid					
Position	P[1]	Red JF	AA[2]	SIFT	Polyphen	PMut	Outcome
384-6	48	R	P	deleterious	benign	-	n/d[3]
			L	deleterious	benign	-	n/d[3]
			Q	deleterious	benign	-	n/d[3]
392-4	51	R	S	tolerated	benign	-	neutral
			G	tolerated	benign	-	neutral
465-7	75	S	T	tolerated	benign	neutral	neutral
			A	tolerated	benign	neutral	neutral
1182-4	202	M	T	tolerated	benign	-	neutral
			A	tolerated	benign	neutral	neutral
1242-4	222	T	A	n/d[3]	possibly damaging	neutral	n/d[3]
			P	deleterious	possibly damaging	-	deleterious

[1] Amino acid site. [2] Alternative alleles. [3] Not determined.

Like the nonsynonymous SNPs in the chicken samples, a multiple sequence alignment with T-Coffee shows that these appeared to be scattered throughout the protein sequence (Figure 4.8). Two nonsynonymous mutations (at sites 630, 633) in amino acids 103 and 104 are adjacent to the polypeptide cleavage point. The protein substitution impacts of Y103F and A104V were predicted to be neutral, suggesting that these variants could have persisted in the population. Y103F was also a high-frequency variant in chicken populations and all sites polymorphic in chicken are not observed in other birds.

Figure 4.8. Multiple sequence T-Coffee alignment of IL1B sites predicted by PAML M8 to be under selection (red) and sites with nonsynonymous SNPs among the chicken samples (blue).



```
Red JF              M A F V P D L D V L E S S S L S E E T F Y G P S C L C L Q K K P R L D S E H T T V D V Q V T V R K G R G A R S F R R A A
Ceylon JF           M A F V P D L D V L E S S S L S E E T F Y G P S C L C L Q K K P R L D S E H T T V D V Q V T V Q K G R G A R S F R R A A
Grey JF             M A F V P D L D V L E S S S L S E E T F Y G P S C L C L Q K K P R L D S E H T T V D V Q V T V R K G R G A R S F R R A A
Chicken             M A F V P D L D V L E S S S L S E E T F Y G P S C L C L Q K K P R L D S E H T T V D V Q V T V R K G R G A R S F R R A A
Bamboo partridge    M A F V P D L D V L E T S S L S E E T F Y G P S C L C L Q K K P R L D S E H S T V G V Q V T V P K G S G A R S F R P A A
Green JF            M A F V P D L D V L E T S S L S E E T F Y G P S C L C L Q K K P R L D S E H N T V G V Q V T V F K G G G A R S F R R A A
Grey francolin      M A F V P D L D V L E T S S L S E E T F Y G P S C L C L Q K K P R L D S E H S T V D V Q V T V L K G S G A R S F R P A A
                    * * * * * * * * * * * : * * * * * * * * * * * * * * * * * * * * * * * * * * . * * . * * * * *   * *   * * * * * *   * *

Red JF              V L V V A M T K L L R R P R S R D F A D S D L S A L L E E V F E P V T F Q R L E S S Y A G A P A F R Y T R S Q S F D I F
Ceylon JF           V L V V A M T K L L R R P R S R D F A D S D L S A L L E E V F E P V T F Q R L E S S Y A G A P T F R Y T R S Q S F D I F
Grey JF             V L V V A M T K L L R R P R S R D F A D S D L S A L L E E V F E P V T F Q R L E S S Y A G A P A F R Y T R S Q S F D I F
Chicken             V L V V A M T K L L R R P R S R D F A D S D L S A L L E E V F E P V T F Q R L E S S S V G A P A F R Y T R S Q S F D I F
Bamboo partridge    V L V V A M T K L L R R P T C R D F A D S D L S A L L E E V F E P V T F Q R L E S S Y A G A P A F R Y T R S Q S F D I F
Green JF            V L V V A M T K L L R R P K C R D F A D S D L S A L L E E V F E P V T F Q R L E S S Y A G A P A F R Y T R S Q S F D I F
Grey francolin      V L V V A M T K L L R R P T R R D F A D S D L S A L L E E V F E P V T F Q R L E S S Y A G A P A F R Y T R S Q S F D I F
                    * * * * * * * * * * * * *     * * * * * * * * * * * * * * * * * * * * * * * * * * *   . * * * : * * * * * * * * * * * *

Red JF              X X X X X X X X X X X X X X X X X X X X X X X X X X X X X R L N I A L Y R P R G P R G S A G T G Q M P V A L G I K G Y K
Ceylon JF           X X X X X X X X X X X X X X X X X X X X X X X X X X X X X R L N I A L Y R P R G P R G S A G T G Q M P V A L G I K G Y K
Grey JF             X X X X X X X X X X X X X X X X X X X X X X X X X X X X X R L N I A L Y R P R G P R G S A G T G Q M P V A L G I K G Y K
Chicken             X X X X X X X X X X X X X X X X X X X X X X X X X X X X X G L N I G L Y G P R G P R G T A G T G Q M P V A L G I K G Y K
Bamboo partridge    X X X X X X X X X X X X X X X X X X X X X X X X X X X X X R L N I A L Y R P R G P R G S A G T G Q M P V A L G I N G Y K
Green JF            X X X X X X X X X X X X X X X X X X X X X X X X X X X X X R L N I A L Y R P R G P R G S A G T G Q M P V A L G I N G Y K
Grey francolin      X X X X X X X X X X X X X X X X X X X X X X X X X X X X X R L N I A L Y R P R G P R G S A G T G Q M P V A L G I N G Y K
                    * * * * * * * * * * * * * * * * * * * * * * * * * * * * *   * * * . * *   * * * * * * : * * * * * * * * * * * * : * * *

Red JF              L Y M S C V M S G T E P T L Q L E E A D V M R D I D S V E L T R F I F Y R L D S P T E G T T R F E S A A F P G W F I C T
Ceylon JF           L Y M S C V M S G T E P T L Q L E E A D V M R D I D S A E L T R F I F Y R L D S P T E G T T R F E S A A F P G W F I C T
Grey JF             L Y M S C V M S G T E P T L Q L E E A D V M R D I D S A E L T R F I F Y R L D S P T E G T T R F E S A A F P G W F I C T
Chicken             L Y M S C V M S G T G P T L Q L E E A D V M R D I D S V E L T R F I F Y R L D S P T E G T T R S E S A A F P G W F I C T
Bamboo partridge    L Y M S C V M S G T E P T L Q L E E A D V T R D I D S V E L T R F I F Y R L D S P A E G T T R F E S A A F P G W F I C T
Green JF            L Y M S C V M S G A E P T L Q L E E A D V T R D I D S V E L T R F I F Y R L D S P A E G T T R F E S A A F P G W F I C T
Grey francolin      L Y M S C V M S G T E P T L Q L E E A D V R R D I D S V E L T R F I F Y R L D S P P E G T T R F E S A A F P G W F I C T
                    * * * * * * * * * :   * * * * * * * * * *   * * * * * . * * * * * * * * * * * * * . * * * * *   * * * * * * * * * * * *

Red JF              S L Q P R X X X X X X X X X X X X X X X X X K L S G R
Ceylon JF           S L Q P R X X X X X X X X X X X X X X X X X K L S G R
Grey JF             S L Q P R X X X X X X X X X X X X X X X X X K L S G R
Chicken             S L Q P R X X X X X X X X X X X X X X X X X K L S G R
Bamboo partridge    S L Q P R X X X X X X X X X X X X X X X X X K L R G P
Green JF            S L Q P R X X X X X X X X X X X X X X X X X K L R G P
Grey francolin      S L Q P R X X X X X X X X X X X X X X X X X K L R G R
                    * * * * * * * * * * * * * * * * * * * * * * * *   *
```

The cleavage point of the pro-peptide is shown between sites 105 and 106 (green). Regions with poor sequence quality are masked out (X).

4.4 Discussion

This was the first study of evolution in functional chicken immune genes in a large set of diverse populations. Two disease-associated genes (IL1B and IFNG) were re-sequenced in seven Asian and African populations and six outgroup samples. Analysis of SNP data, summary statistics and coalescent simulations suggested that diversity within the two genes was different and particularly high at IL1B. Tests of neutrality indicated the presence of balancing selection at this gene and PAML analysis supported the possibility of adaptive processes. Confounding factors for determining selection included recombination, which was elevated at IL1B.

4.4.1 Elevated and contrasting diversity:

IL1B and IFNG displayed differences in diversity as a result of the different population histories. The International Chicken Polymorphism Map Consortium (2004) found π levels of the same scale between red JF and domestic chickens (5.36 per kb on average) as described for IL1B (4.92). Most diversity statistics were unusually low at IFNG, where π was 3.18 per kb. The contrast in diversity at these two genes was visually apparent in haplotype distribution in the network diagrams. IL1B had a higher frequency of rare alleles, higher haplotype diversity and a higher concentration of SNPs.

The AMOVA results, pairwise F_{ST} trees and Mantel permutation tests illustrated the absence of either strong population structure or of an association of population distribution with geography at either gene. The high diversity within populations was consistent with the idea that chickens did not endure a substantial bottleneck during domestication (Ellegren 2005), although the pattern of this diversity may be affected by other events, such as introgression of wild JF.

One key inference from the time depths estimated from the pairwise difference plots was that even if large population size increases and recombination have exaggerated the extent of measured variation, most SNPs present in domestic and wild lines pre-date domestication (International Chicken Polymorphism Map Consortium 2004). This may be still functionally interesting and can illuminate ancient population history. For example, the incidence of the Mx susceptibility allele is higher in broilers than layers and this was present in ancestral strains (Balkissoon et al. 2007). This could be compounded by a range expansion or an increase in population size, which can both mimic directional selection (Excoffier 2004).

The lack of differentiation and geographical structuring among the chicken populations and continents may be an artefact of the high portability and tradability of the chicken during its history since domestication. Results from Liu et al. (2006) showed geographic structuring in chickens, suggesting chicken autosomal DNA may have different population histories to that of mtDNA.

4.4.2 Differing levels of recombination and GC content:

The chicken genome is noted for its high rate of chromosome-specific recombination, particularly at microchromosomes (Ellegren 2005). As evident in high DnaSP R and LDhat ρ values, IL1B had an exceptionally elevated rate of recombination, a level not accounted for by being on microchromosome 22. This extreme recombination rate is likely to have had profound effects on distribution of diversity and so would have altered the signature of selection observed. Though consistent with its extensive conservation, the low recombination rate at IFNG further highlights the contrast between the two genes.

The significantly negative Fu's F_S at IL1B was due to an excess of rare alleles and is characteristic of directional selection (Akey et al. 2004). Recombination can lead to an excess of rare variants by increasing the numbers of SNPs and reducing the genetic distance between haplotypes (Tajima & Mukai 1990, Tajima 1993, McVean et al. 2002, Pennings & Hermisson 2006) and may mimic the effects of high positive selection (Reed & Tishkoff 2006)

GC content is an indicator of the recombination rate (Duret & Arndt 2008), and correlates with recombination hotspots (Gordon et al. 2007, Buard & de Massy 2007). The elevated GC content at IL1B (0.65) was above the genome average (0.42; International Chicken Genome Sequencing Consortium 2004) and the chromosomal average (0.43; Gao & Zhang 2006). Given that functional sequences are not normally GC rich (Galtier & Duret 2007), this was a unique feature of this gene. GC content at

IFNG (0.43) was not substantially different from that of the genome or chromosome 1 (0.40), where it is located.

The statistic R, which measures crossing over and gene conversion (Hudson 1987), was much higher at IL1B compared to IFNG than the relative difference between them for R_M, the minimum number crossing over events (Hudson & Kaplan 1985). Given this and the higher number of gene conversion events at IL1B compared to IFNG, excess recombination at IL1B could be due, in part, to biased gene conversion (BGC), a characteristic of GC-rich regions (Ellegren 2007). BGC is likely to have played a significant role in shaping the chicken genome, particularly at certain immune genes (Das et al. 2009). The GC isochores resulting from BGC may be subject to selective pressures (Webster et al. 2006).

4.4.3 Evidence for selection:

A number of tests of neutrality at IL1B generated extreme values. IL1B had a significant excess of intermediate alleles, which suggests the presence of balancing selection. Further evidence for this was found in the coalescent simulation results, which show that the Tajima's *D* value for IL1B was much more positive than expected across a set of demographic scenarios.

Fay and Wu's *H* was significantly negative for IL1B, indicating a preponderance of derived alleles (Fay and Wu 2000). The presence of a highly negative *H* value in the IL1B coding region but not in the non-coding region was evidence that the *H* value may be a result of selection at coding sites. Fu and Li's *D* and *F* indicated a deficit of singletons at IL1B, suggesting it may be conserved to some degree, particularly in the Asian set of samples (Fu & Li 1993).

Interestingly, the values for Tajima's *D* and Fay and Wu's *H* are more extreme in Africa than Asia. Signatures of selection relating to new environments may be stronger in Africa than in Asia because chickens were first domesticated in Asia, and thus it was possible that African chickens share a more recent history.

More evidence for the presence of adaptation at IL1B lies in the codeml ω-ratio tests. Lineage-specific tests showed that the chicken branch had a ω value higher than the other birds. Site-specific test results produced five sites that were candidates for selection: the 48th, 51st, 75th, 202nd and 222nd amino acids. These were segregating between chicken and the other bird species: Ceylon JF, green JF, bamboo partridge and grey francolin. Thus, it was possible selective forces acted on these sites during the evolution of the *Gallus* lineage, or that selection constraints have become more relaxed among avian species. The first three sites lie in the IL1B precursor portion that is cleaved between sites 105 and 106 to produce the active mature polypeptide of 162 amino acids (Weining et al. 1998, Gyorfy et al. 2003), suggesting their functional role may differ from that of the other two amino acids (Figure 4.5).

The pairwise difference plots of IL1B for all sites and for coding sites showed significant disparities: the former suggested an ancient expansion of diversity or population size, whereas the latter was consistent with a neutral increase of diversity – indicating conservation of the coding sequence. This lack of CDS variation may indicate the part of the balanced signal is partially due to the complicated demographic history of the chicken. Though, directional selection and purifying selection would have the effect of reducing diversity (Harris and Meyer 2006).

Most sites in key functional immune genes, such as the two examined here, can be expected to be conserved, with only limited numbers of sites under positive selection (Yang et al. 2001). There was evidence for conservation in both genes, particularly in the coding regions. This was most apparent at IFNG, where there was just one synonymous SNP among 68 detected in the 140 genotypes. Furthermore, the only difference between the red JF genome sequence and the most numerous haplotype at IFNG was one SNP. In sharp contrast, IL1B had a large number of coding SNPs among the chicken and outgroup samples and a substantial amount of these are nonsynonymous. This contrast indicates stronger preservation of functional sequence at IFNG, most likely by purifying selection.

Interestingly, the pattern of a high number of nonsynonymous changes, a high ω value and evidence for selection seen here for the IL1B gene was a trend observed not only in the Mx gene (Hou et al. 2007, Berlin et al. 2008), but also in the MHC-B gene (Worley et al. 2008). This could represent a pattern of chicken immune system genes that maintain diversity in order to respond to a wider variety of pathogens.

4.5 Conclusion

This study shows how comprehensive sampling can reveal distinctive patterns of diversity at two disease-associated chicken genes. IFNG had low diversity among the chicken and outgroup samples, and showed a high degree of coding sequence conservation; both of these observations are evidence for its function as a pivotal regulator of the immune system. IL1B diversity possessed properties symptomatic of both balancing selection and recombination, yielding high numbers of diverse alleles. This could be due to challenges driven by new diseases in different environments, or it could represent an indication of multiple functions for IL1B in the chicken, as it has in mammals.

It was already established that the chicken was domesticated in multiple locations (Liu et al. 2006) and that wild red JF and domestic village strains may be closely related (Kanginakudru et al. 2008). Networks shown here indicate that red JF was the closest outgroup and therefore the most likely ancestor for diversity at each gene. This was in contrast to the yellow skin gene (Eriksson et al. 2008), which appears to originate in an introgression from grey JF. This seems to exclude the introduction of exogenous variants in a gene with high diversity and balancing selection signals in IL1B. However, this may be influenced by repeated introgression and interbreeding of domestic populations with wild red JF.

Acknowledgement

Some of the perl scripts (PhaseIn.pl, SeqBuild.pl, hashadd.pl and maskseqs2.pl) and a Microsoft Excel application (by Stephen Park) used to parse the Phred-Phrap-Polyphred-Consed output were originally written by Ruth Freeman, Brian McEvoy and David Lynn.

Publication

This chapter formed the basis for a publication in Immunogenetics 61(4):303-314 (erratum in 61(5):421–422) in 2009 entitled "Contrasting evolution of diversity at two disease-associated chicken genes". The authors are: Downing T, Lynn DJ, Connell S, Lloyd AT, Bhuiyan AKFH, Silva P, Naqvi A, Sanfo R, Sow RS, Podisi B, O'Farrelly C, Hanotte O and Bradley DG.

CHAPTER 5

Bioinformatic discovery and population-level validation of selection at the chicken interleukin-4 receptor alpha-chain gene

5.1 Introduction

New large-scale sequencing projects in several avian species, for instance the zebra finch genome project (http://songbirdgenome.org), now allow the genome-wide comparative analysis of avian genes and the detection of selection on a more comprehensive scale. The estimated 100 million years of divergence between zebra finch and the chicken permits the robust evaluation of functionally relevant evolutionary change (Kaiser et al. 2007). For many chicken genes, the protein-level identity with mammalian species like human and mouse is too low to permit effective analysis of functional variation due to the long time since species divergence (International Chicken Genome Sequencing Consortium 2004).

Approximately 20% amino acid changes between chicken and zebra finch have been fixed by positive selection (Axelsson et al. 2009), so by comparing CDS between these (and other) birds, chicken genes with signals suggestive of adaptation can be identified. Previous examinations of chicken genes defined by mammalian orthologs suggested immune genes in particular undergo a more frequent rate of change at the protein level (International Chicken Genome Sequencing Consortium 2004), so testing for substitutions that are adaptive in this set of genes may be an successful strategy for identifying variability relevant to disease.

5.1.1 IL4RA gene functions:

This chapter reports that the chicken interleukin receptor 4 alpha chain gene (IL4RA) showed a relative excess of nonsynonymous substitutions and may be subject to selection. Chicken IL4RA is associated with disease: for example, its expression is downregulated by avian influenza virus during infection (Xing et al. 2008).

The human ortholog of this gene encodes a type 1 transmembrane receptor for IL4 and IL13, both of which are key immune system cytokines that initiate signalling pathways in the inflammatory response to infection (Shirakawa et al. 2000). IL4RA may also form an interleukin receptor with a gamma-c (γ_c) receptor chain – γ_c can bind other cytokines (Hershey et al. 1997). In humans, IL4RA regulates IgE production by B cells and differentiation of T_H2 cells (Wu et al. 2001, Liu et al. 2004). Notably, replacement substitutions at human IL4RA are associated with the onset of atopy through the over-activation of inflammation (Hershey et al. 1997); this suggests such variants would have historically been mildly deleterious.

Interestingly, an alternate form of the human protein with a different C-terminus can be produced by translation or proteolysis of the dominant transmembrane variant: this novel protein can inhibit cell proliferation driven by IL4 and IL5 expression by T cells (Bergin et al. 2006). In humans, the extracellular domain is made up of exons 3 to 7, the transmembrane domain from exon 9, and the intracellular domain exons 10 to 12; the shorter, soluble IL4RA version has no transmembrane or intracellular domains (Kruse et al. 1999).

The IL4RA gene was resequenced in 70 Asian and African village chickens, 20 commercial broilers, and in six closely related species: red, grey, Ceylon and green JF, bamboo partridge and grey francolin. High allelic variation at this gene appeared to be balanced at two nonsynonymous SNP sites in particular. Although this may enhance immune system variability in response to challenges by pathogens, a consequence of the complex domestication history of the chicken is that introgression, multiple origins and migration are likely to have altered the pattern of diversity at this locus, complicating selection signatures.

5.2 Methods

5.2.1 Identification of putative alignments of chicken genes

As the most extensively sequenced other bird species, zebra finch genes were compared with the chicken genome. With adequate levels of turkey (*Meleagris gallopavo*) sequence present in GenBank, this species was used as an additional contrast. For example, at the time of writing, the mallard duck (*Anas Platyrhynchos*) was the bird species with the fourth highest number of GenBank sequences but still had over four times fewer sequences than turkey.

Chicken Refseq protein sequences (19,661), zebra finch ESTs and mRNAs (67,671), and turkey ESTs and mRNAs (16,032), as well as zebra finch (264) and turkey (39) BAC sequences were downloaded from GenBank. These were cleaned of vector contaminants using SeqClean and repetitive sequences were masked using RepeatMasker. Tgicl (Pertea et al. 2003) was used to cluster the zebra finch and turkey sequences separately with a minimum length of 100 bases and identity of 96% or more for overlapping regions into 9,716 zebra finch and 1,810 turkey consensus contigs (Figure 5.1).

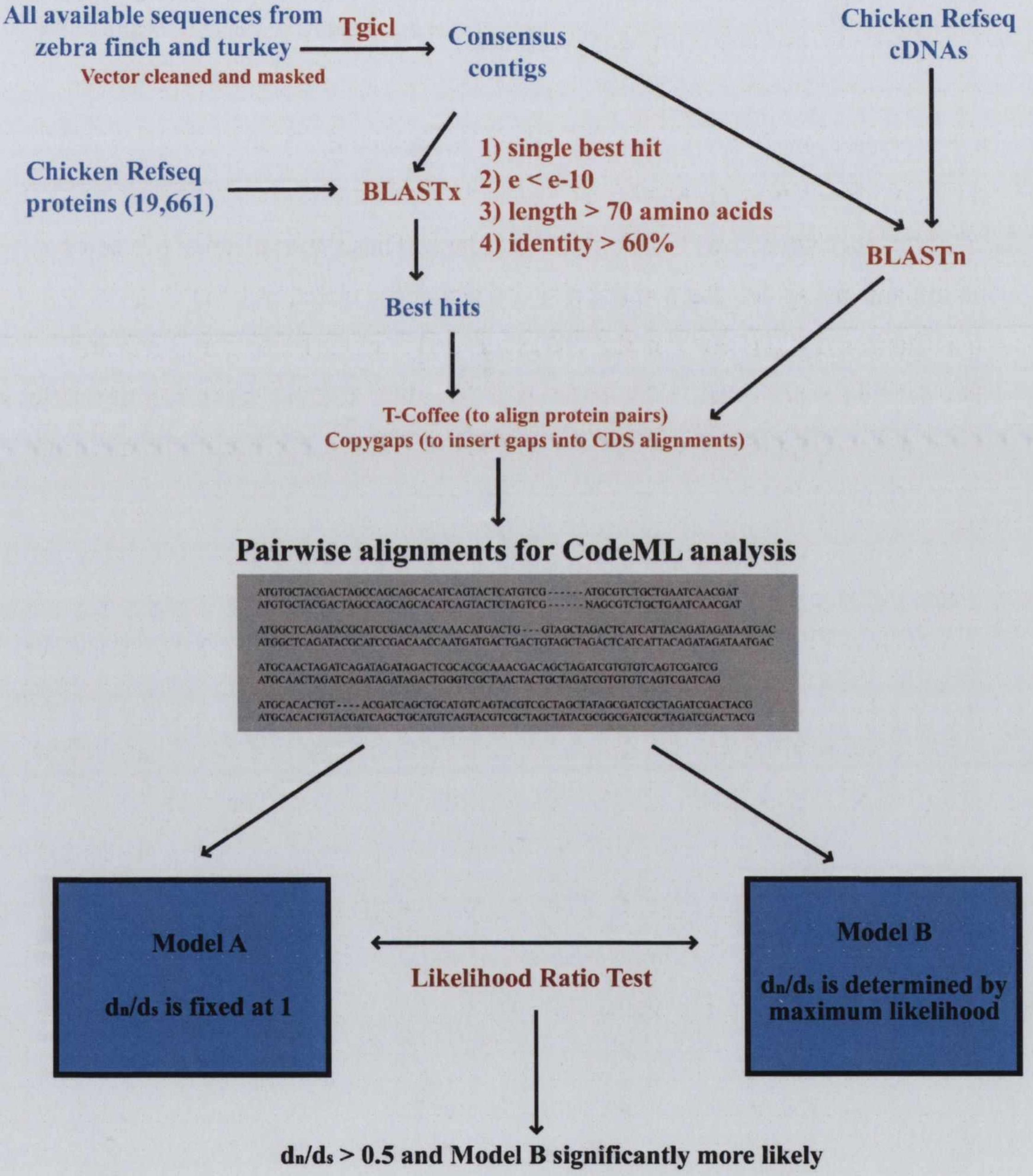
The zebra finch and turkey contigs were searched against the chicken protein sequences using Blastx (Gish & States 1993), with an E value ≤ e-10 separating best hits for each protein from paralogous sequences. The best-hit protein pairs identified in the Blastx search were aligned with T-Coffee (Notredame et al. 2000) using perl scripts (ckzfNEWblastx.pl, BLASTX.pl and hitParserZF.pl in Appendix A). Alignments of length < 70 amino acids or sequence identity < 60% were discarded to remove short or spurious sequences. These protein alignments were then used as templates to generate 3,653 chicken-zebra finch and 1,139 chicken-turkey pairwise coding sequence (CDS) alignments in the correct reading frames and with gaps inserted where needed. These were used in subsequent analyses. The chicken-zebra finch pairs were also used for analysis in Chapter 2.

5.2.2 Identifying candidate genes subject to selection:

Pairwise d_N/d_S (ω) was calculated for each CDS alignment using the codeml implementation of the PAML 3.15 package (see Chapter 2 for details; Yang 1997). ω

was compared by maximum likelihood under two different models: a neutral model (Model A) where ω was fixed = 1, and a model where ω was free to vary (Model B). These models were compared using a LRT to determine if the variable model was significantly more likely (Yang 1997).

Figure 5.1. Procedural pipeline for determining orthologous alignments of chicken genes that are candidates for undergoing selection.


All available sequences from zebra finch and turkey
Vector cleaned and masked
Tgicl
Consensus contigs
Chicken Refseq cDNAs
Chicken Refseq proteins (19,661)
BLASTx
1) single best hit
2) e < e-10
3) length > 70 amino acids
4) identity > 60%
BLASTn
Best hits
T-Coffee (to align protein pairs)
Copygaps (to insert gaps into CDS alignments)
Pairwise alignments for CodeML analysis
Model A
d_n/d_s is fixed at 1
Likelihood Ratio Test
Model B
d_n/d_s is determined by maximum likelihood
$d_n/d_s > 0.5$ and Model B significantly more likely

```
ATGTGCTACGACTAGCCAGCAGCACATCAGTACTCATGTCG------ATGCGTCTGCTGAATCAACGAT
ATGTGCTACGACTAGCCAGCAGCACATCAGTACTCTAGTCG------NAGCGTCTGCTGAATCAACGAT

ATGGCTCAGATACGCATCCGACAACCAAACATGACTG---GTAGCTAGACTCATCATTACAGATAGATAATGAC
ATGGCTCAGATACGCATCCGACAACCAATGATGACTGACTGTAGCTAGACTCATCATTACAGATAGATAATGAC

ATGCAACTAGATCAGATAGATAGACTCGCACGCAAACGACAGCTAGATCGTGTGTCAGTCGATCG
ATGCAACTAGATCAGATAGATAGACTGGGTCGCTAACTACTGCTAGATCGTGTGTCAGTCGATCAG

ATGCACACTGT----ACGATCAGCTGCATGTCAGTACGTCGCTAGCTATAGCGATCGCTAGATCGACTACG
ATGCACACTGTACGATCAGCTGCATGTCAGTACGTCGCTAGCTATACGCGGCGATCGCTAGATCGACTACG
```

Programs used and their parameters are in red. Datasets are in blue. Pairwise alignments with PAML are in black.

As a consequence of this conservative strategy of calculating ω across the entire gene length, genes may be discounted when the signal of directional selection is focused on specific regions or domains, and would thus be obscured by purifying selection operating on the majority of the gene (Sawyer et al. 2005). Many genes known to be subject to positive selection have $0.5 < \omega < 1$ (Swanson et al. 2004), so using a lower cut-off point than $\omega > 1$ to identify candidate genes that may be subject to selection can be effective. Accordingly, chicken-zebra finch alignments with $\omega > 0.5$ where the variable model was significantly favoured ($p < 0.05$) were identified. The annotation associated with the best human orthologs from the Panther database (Thomas et al. 2003) was used to identify the function of chicken genes with relevance to the immune system.

The chicken IL4RA mRNA sequence (Refseq ID: XM_414885) was initially determined by Boardman et al. (GenBank accession: CR407301) and Caldwell et al. (2004). This sequence aligned as a best hit to 2 clustered zebra finch sequences, DQ213788 and DQ213787 (Wada et al. 2006). Situated on chromosome 14, the 5' end of IL4RA is just 150 bp from a transcribed element (NSMCE1; Caldwell et al. 2005). The IL21 receptor is near the 3' end of IL4RA and an IL9R precursor homolog lies close to the IL21R as well.

5.2.3 Sample collection:

A total of 90 chicken samples were acquired: 70 village birds from Asia and Africa (International Livestock Research Institute, Kenya) and 20 commercial broilers (Manor Farms, Co. Monaghan, Ireland). The commercial birds were composed of 10 Ross breed chickens from Ireland and 10 Hubbard Flex from France. The Asian and African samples were the same as in Chapter 4. One sample for each of 6 outgroup species were also sequenced – again, as per Chapter 4: bamboo partridge, grey francolin, and green, grey, Ceylon and red JF. DNA was isolated from the samples using a phenol-chloroform extraction following a proteinase K digestion.

5.2.4 Resequencing strategy:

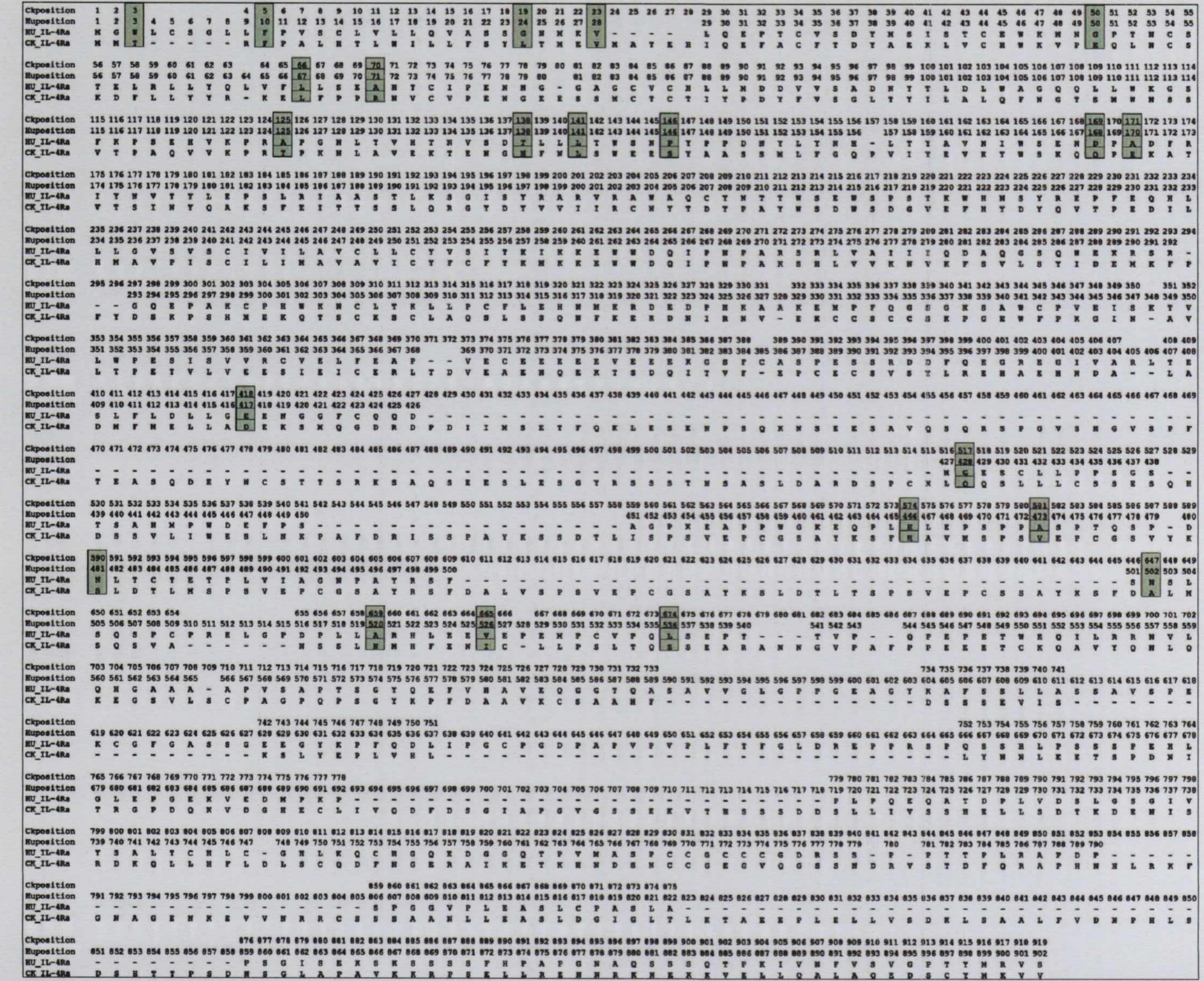
The UCSC, GenBank and Ensembl databases were used to investigate the gene's structure. At the time of analysis, a portion of the chicken IL4RA region was not displayed on these browsers, so the reference assembly (NC_006101) and reference contig (NW_001471454) were aligned with the IL4RA mRNA sequence from GenBank (XM_414885) using T-Coffee (Notredame et al. 2000) to determine potential coding regions. A further T-Coffee alignment of the human and chicken IL4RA protein sequences identified chicken regions orthologous to variable regions in humans (Figure 5.2): according to the Uniprot (www.uniprot.org) entry for human IL4RA (Uniprot: P24394), most variation is in the extracellular and cytoplasmic domains. Genscan was used to corroborate the predicted gene structure (Figure 5.3; http://genes.mit.edu/GENSCAN.html).

PCR primers were designed using Primer3 according to the parameters listed in Chapter 4 and were constructed by VHBio. The details of the primer sequences and optimal parameters for their usage are in Table 5.1. Each amplicon was amplified according to the PCR cycle setup (Table 5.2): eight were successfully amplified for all 96 samples. The forward and reverse PCR product sequences were determined by Agowa.

Table 5.1. Sets of primer pair sequences and their associated optimal PCR parameters.

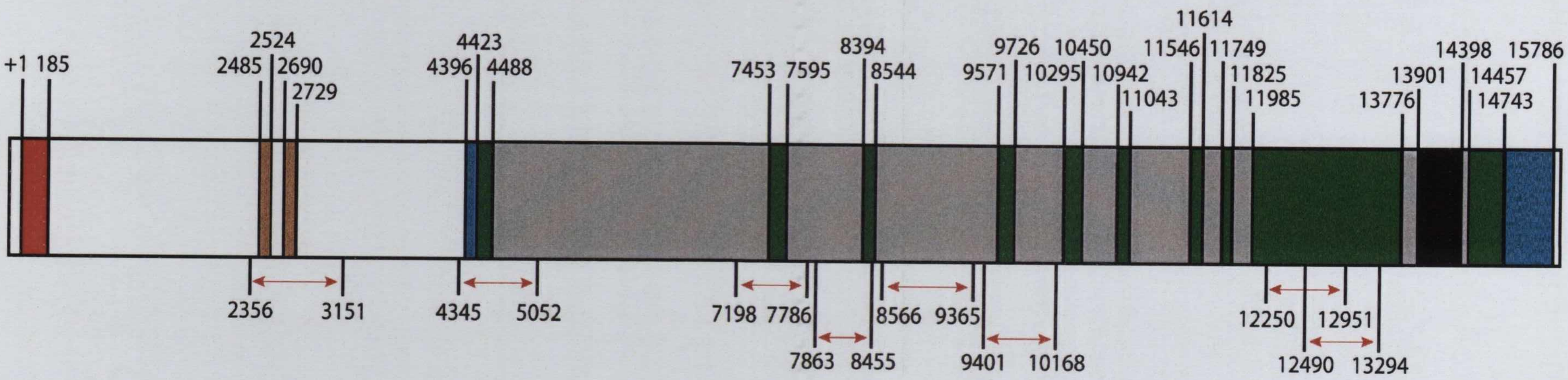
Amplicon	Size (bp)	Orientation	T_M (°C)	[MgCl] (mM)	Primer Sequences
1	903	Forward	56	15	GGTTAGGTTGCAAGGTTTTGTC
		Reverse			CCAGCCCTTAAGATTTCATGTC
2	799	Forward	60	20	GAATCCTAACATCCAGCAAAGC
		Reverse			AGTGAAGAACACACACCACCAC
3	684	Forward	56	20	CAGGAAAAATCCCAACTGAAAG
		Reverse			GCACTACTTGGCAAACACTCTG
4	708	Forward	61	15	CAGAGTGTTTGCCAAGTAGTGC
		Reverse			ACATACTGGTGCCATTGAACTG
5	943	Forward	57	25	ACAGTTCAATGGCACCAGTATG
		Reverse			TTCAGGCCTTCTCACTAAGCTC
6	867	Forward	58	20	GCAGTGCTTGTTGATGAATACC
		Reverse			TTAGATGCCAACTGTGTTGTCC
7	970	Forward	60	20	AATGCAGTTTTAACCCCTGAGA
		Reverse			GGGTTAAAGACGGTAACAAGCA
8	906	Forward	62	20	ACAATTGCAGTACAACCAGCAG
		Reverse			TCAAACACTCATGGCCATCTAC

Figure 5.2. An alignment of chicken and human IL4RA protein sequences.



The consensus human IL4RA sequence isoform a (GenBank accession number NP_000409) and the consensus chicken sequence (XP_414885) were aligned with T-Coffee. The sites marked green were subsequently found to be candidates for selection according to PAML M8 BEB results. Sites marked green and in red letters indicate those subsequently observed as segregating in chicken populations and/or with differences between the chicken and the red JF sequences.

Figure 5.3. Gene structure of IL4RA.



Exons are shown in green, introns in grey and amplicon regions by the red arrows. The UTRs are shown in blue, the leader sequence in red, unknown regions in black and promoter sequence is shown in beige. The numbers shown represent the base positions in relation to the GenBank entries for the mRNA and CDS.

Table 5.2. PCR cycle program for each primer pair.

Step	Temp. (°C)	Duration
1	95	15 mins
2	95	0.5 min
3	T_M [1]	0.75 min
4	72	1 min
5	72	15 mins

[1] Annealing temperature as listed in Table 5.1. Steps 2 to 4 were repeated 33 times in sequence.

5.2.5 Sequence assembly:

Sequencing reads were assembled into contigs using the Phred-Phrap-Consed-Polyphred pipeline programs Phrap v0.990319 and Phred v0.020425.c (Ewing & Green 1998, Ewing et al. 1998). For complete details of base calling, SNP detection and sequence assembly see Chapter 4, though elements were improved from Chapter 4: firstly, only bases with base quality scores (S) > 20 were included in the analysis, so all bases had at least a 99.0% probability of being correct: most had $S \geq 40$ (99.99%). And secondly, only SNPs in polyphred rank 1 were called for the outgroup samples.

A list of the genotypes for each sample was collated and PHASE version 2.1.1 (Stephens et al. 2001) was used to infer missing haplotypes. These assigned haplotypes were cross-referenced with haplotypes generated by Arlequin (Schneider et al. 2000) to ensure consistency – both were identical. Any sequence sites with inadequate coverage across populations or continents, which had sub-standard base quality scores, or had insufficient coverage for either forward or reverse sequences, were removed using Perl scripts – leaving a total of 5,298 bp for further analysis. Coding sequence regions were corroborated using MGalign version 3.1 (Lee et al. 2003). The sequences were exported to Mega (version 4.0.2, Tamura et al. 2007) to convert the data to formats useable by other software packages.

5.2.6 Data analysis:

DnaSP 4.0 (Rozas & Rozas 1999, Rozas et al. 2003) was used to analyse the polymorphic characteristics of the data and to perform a series of population genetic analyses, the calculation details of which are discussed in Chapter 4. The numbers and types of SNPs were assessed. Nucleotide diversity was measured using π, the average number of nucleotide differences between sequences pairs (Tajima 1983). The haplotype diversity (*Hd*, the number and frequency of haplotypes in the sample, Equation 2; Depaulis & Veuille 1998), the number of haplotypes, Z_{nS} (Kelly 1997, Equation 3) and $\theta_W = 4N_e\mu$ (Watterson 1975, Equation 4) were determined. The four gamete test for the minimum number of recombination events (R_M; Hudson & Kaplan 1985, Equation 5) and *R* (the degree of recombination; Hudson 1987, Equation 6) were calculated, as was the GC content.

A set of summary statistics were used to identify departures from neutrality using simulations: Fu and Li's *D* and *F* (Fu & Li 1993, Equations 8 and 9), Tajima's *D* (Tajima 1989, Equation 7), Fu's F_s (Fu 1993, Equation 11) and Fay and Wu's *H* (Fay & Wu 2000, Equation 10). For details on their calculations and the coalescent simulations using DnaSP, see Chapter 4. These simulations generated empirical distributions with which the statistical values were compared to determine the extent of their deviation from neutrality. It is an indication of non-neutral evolution if the observed values lie at the extremes of the distribution.

Median-joining haplotype networks were constructed using Network version 4.2.0.1 (Bandelt et al. 1999). AMOVA tests (Excoffier et al. 1992) were conducted using Arlequin (Schneider et al. 2000) with 1,000 permutations. See Chapter 4 for details on the utility of the test.

Predictions to estimate the extent of functional impact for each radical substitution were conducted using PMut (Ferrer-Costa et al. 2005) – more details on the test are in Chapter 4. In some cases, the program did not have sufficient confidence in the results due to the high protein sequence divergence between chicken and the species with which it was compared. In such cases, the prediction outcomes were classed as not determined.

The McDonald-Kreitman tests (McDonald & Kreitman 1991) were implemented with DnaSP to examine the rates of evolution within a species (chicken here) to that between species (between chicken and outgroup genotypes) at two categories of sites. The relative ratios of fixed nonsynonymous (D_N) and silent (D_L) substitutions and polymorphic nonsynonymous (P_N) and silent (P_L) changes are evaluated as D_N/D_L and P_N/P_L. Silent sites include both noncoding and synonymous sites. The test calibrates the rates of nonsynonymous site change for what is assumed to be a neutral rate at silent sites. If $D_N/D_S > P_N/P_S$ or $D_N/D_L < P_N/P_L$, based on a one-tailed Fisher's Exact Test, it can indicate the presence of non-neutral evolution (McDonald & Kreitman 1991). If $D_N/D_L > P_N/P_L$, it is more consistent with purifying selection on the ancestral interspecies branch (Eyre-Walker 2002): this would likely require synonymous rather than silent sites in order to detect positive selective in a more stringent manner because the probable lack of selective constraint on coding compared to noncoding sites may bias the test, so using synonymous sites would reduce this inaccuracy, given that differing rates of background selection would mimic the effects of positive selection (Eyre-Walker 2006).

5.2.7 Selection at IL4RA among avian species:

To investigate for evidence of selection in IL4RA between chicken and each of the six outgroups, CDS alignments were generated and ω was determined under a variety of models using codeml (Yang 1997). For this analysis, a chicken sequence from the most numerous haplotype was used (FJ542575). Although the chicken coding haplotypes observed at IL4RA were diverse, substituting this for other chicken genotypes yielded no significant changes to results, except at certain sites with model M8 for a divergent sample (FJ542675).

The free-ratio (M1) model was used to calculate tree branch lengths and ω for each species lineage in the sample. To identify specific codon sites with evidence of selection, site-specific models estimated ω for each site across the whole sequence by using a random sites model under BEB (for details see Chapter 4; Yang 1997, Nielsen & Yang 1998, Yang et al. 2005). A LRT was conducted between the paired neutral and variable models (neutral M1a vs variable M2a; neutral M7 vs variable M8). BEB determined the posterior Bayesian probability of ω for each amino acid site: a significantly high posterior probability for this variable ω class suggests that a particular site is under selection, if $\omega > 1$ and M8 (or M2a) is significantly favoured by the LRT (Yang et al. 2005). Candidate positively selected sites from M8 were examined using PMut to assess the functional impact for each nonsynonymous substitution.

5.2.8 Identification of IL4RA in the zebra finch

Searching the zebra finch genome sequence (July 2008 assembly) with the chicken IL4RA protein sequence (XP_414885) and the translated versions of zebra finch sequences (DQ213788, DQ231787) using tBlastn (Altschul et al. 1990) identified the IL4RA gene on zebra finch chromosome 14. Alignments of known bird IL4RA gene and protein sequences and the candidate zebra finch region on chr14 using T-Coffee (Notredame et al. 2000) and the tBlastn data yielded a large portion of the translated zebra finch IL4RA coding sequence.

5.3 Results

5.3.1 Pairwise comparisons of chicken and zebra finch genes:

A set of 3,653 chicken-zebra finch and 1,139 chicken-turkey CDS pairwise alignments were examined for candidate genes potentially subject to directional selection. After visual screening, 12 valid chicken candidate genes were observed with $\omega > 0.5$ where the variable model was significantly favoured (Table 5.3, Figure 5.4).

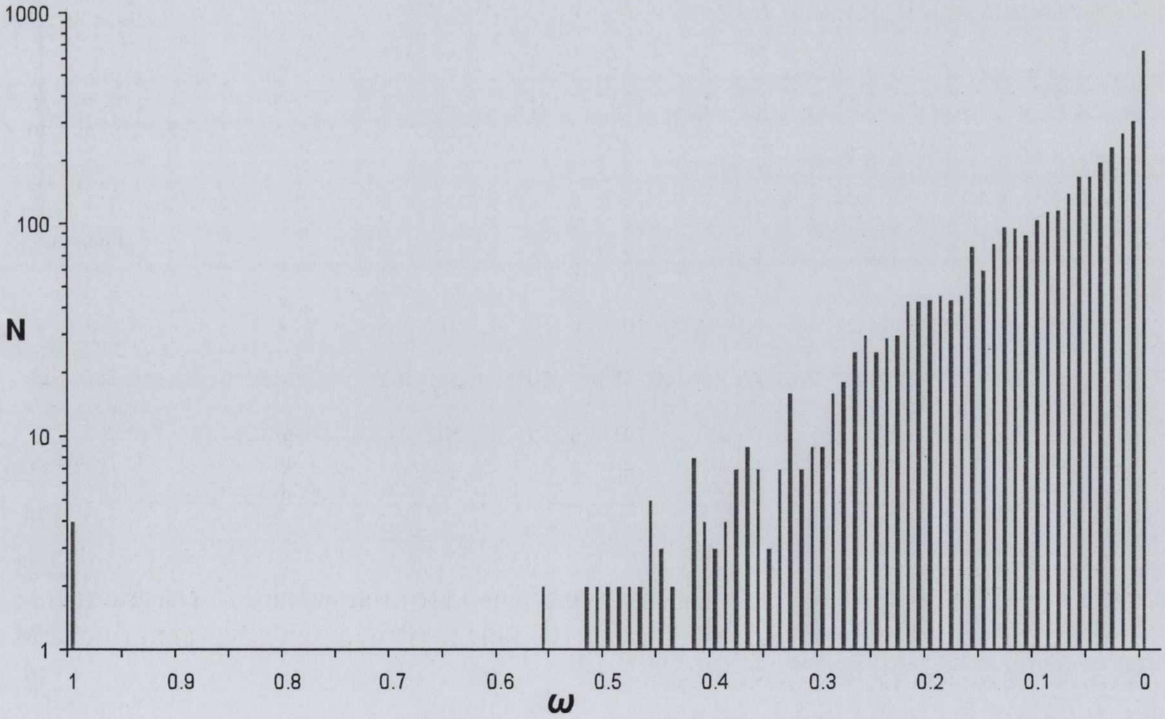
Table 5.3. Pairwise comparison details and functions for chicken sequences with $\omega >$ 0.5 and $p < 0.05$.

Chicken Refseq accession number	$\omega = d_N/d_S$	$2\Delta ML$	P value	d_N	d_S	Chicken gene name	Orthologous human gene function	Chicken GenBank description
XM_420574	3.0968	5.668	0.0173	3.741	1.208	LOC422614	Signalling	Protein phosphatase 1K (PP2C domain containing)
XM_414705	0.5373	4.334	0.0374	0.225	0.418	NDUFB6	Structure	NADH dehydrogenase (ubiquinone) 1 β subcomplex 6
XM_419473	0.5162	4.776	0.0289	0.201	0.390	SLC4A1AP	Signalling	Solute carrier family 4 (anion exchanger), member 1, adaptor
NM_001012594	0.5127	4.430	0.0353	0.157	0.306	GORASP2	Structure	Golgi reassembly stacking protein 2
NM_001031332	0.5511	8.708	0.0032	0.254	0.461	GTSE1	Apoptosis	G-2 & S-phase expressed 1
XM_414885 [2]	0.5098	9.896	0.0017	0.179	0.351	LOC416585	Immunity	IL4R α-chain
NM_001030626	0.5665	6.796	0.0091	0.426	0.751	PIAS2	Immunity	Protein inhibitor of activated STAT, 2
XM_417014	0.5666	4.006	0.0453	0.115	0.202	LOC418820	Immunity	Progesterone-induced blocking factor 1
XM_420836	4.0730	6.172	0.0130	4.027	0.989	LOC422894	-	-
XM_001234647	0.5082	7.616	0.0058	0.544	1.070	LOC771361	-	-
XM_001234647	0.5215	3.988	0.0458	0.197	0.377	LOC771361	-	-
XM_418660 [3,4]	0.5700	17.036	<0.0001	0.055	0.096	LOC420559	-	KIAA2005, sterile α motif domain containing 9
NM_205033 [3]	0.5983	7.306	0.007	0.191	0.319	NES	Apoptosis, neuro-development	Nestin
NM_205033 [3]	0.6108	6.350	0.012	0.120	0.196	NES		

[1] Twice the difference of the maximum likelihood values of the variable model minus the fixed model. [2] IL4RA, the chicken gene selected for resequencing. [3] These were identified through comparisons with turkey, and the remainder with the zebra finch. [4] Aligned with turkey BAC sequence.

The most represented functional category among the 12 candidate genes was related to immunity. Three genes have roles in the immune response: IL4RA, protein inhibitor of activated STAT 2 (Pias2) and progesterone-induced blocking factor 1 (Pibf). Other functional categories included apoptosis: G-2 and S-phase expressed 1 stimulates p53 localisation and controls DNA damage-induced apoptosis in humans (Monte et al. 2003). Also represented was nestin, a type IV intermediate filament protein that is expressed during cranial development in humans to divide nervous system cells into types (Lendahl et al. 1990). Signalling genes are listed as a phosphatase and an anion exchanger – the latter is an anion exchange adaptor in humans. The human version of the intracellular structure gene NADH DH 1β 6 encodes two transcripts that create a mitochrondrial protein. GORASP2 is required for golgi fragmentation during apoptosis in humans. Functions for two genes were unknown. Two of the genes with $\omega > 1$ were not valid coding sequences; the other two were a phosphatase (PPM1K) and an unannotated sequence (XM_420836). From these genes, IL4RA was selected for further analysis because of its critical function in the immune response, including an implicated role in the anti-viral response (Xing et al. 2008).

Figure 5.4. The numbers of genes (N) in classes of ω values from pairwise alignments of chicken-zebra finch gene sets where the variable model was favoured (p<0.05).



The y-axis is on a logarithmic scale. The ω values on the x-axis are classes into groups of 0.01, with the exception of values greater than 1, which are classed as 0.99-1.00 (with $N = 4$).

The IL4RA mRNA sequence (XM_414885) aligned as a best hit to two clustered zebra finch mRNAs (DQ213788 and DQ213787) in contig CL6154Contig1 with a Blastx score of 339 and an E-value of 2e-92. A LRT of the variable and fixed model pairwise comparison log-likelihoods showed that the variable model ($\omega = 0.5098$) was significantly more likely than the neutral model ($\omega = 1$; -1422.79 for variable vs -1427.74 for fixed, p = 0.0017).

5.3.2 Exploring interspecies selection at IL4Rα:

IL4RA was resequenced in seven closely related bird species: chicken, red JF, grey JF, Ceylon JF, green JF, grey francolin and bamboo partridge. An excess of nonsynonymous compared to synonymous substitutions was observed in all birds except red JF (Table 5.4).

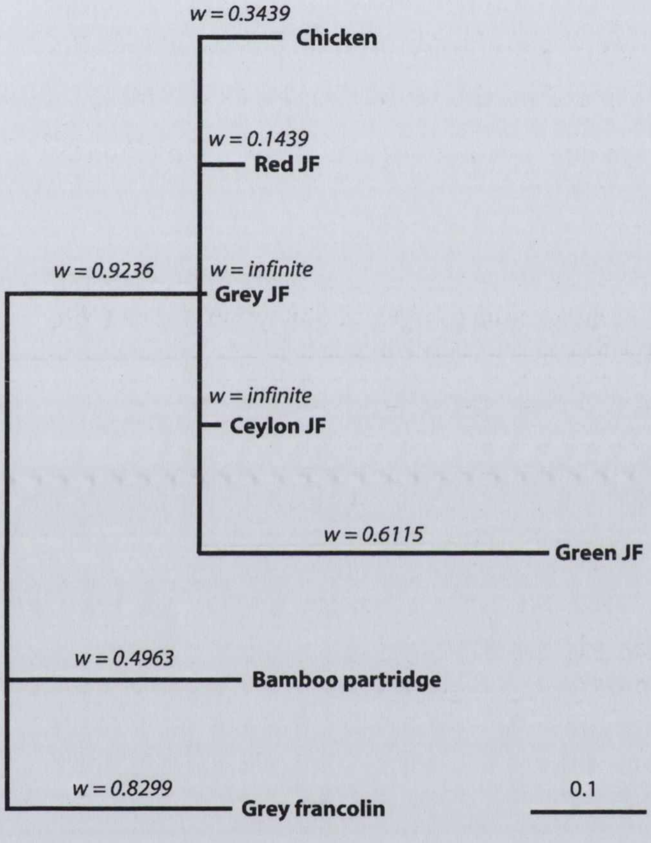
Table 5.4. Estimated distribution of synonymous ($S.d_S$) and nonsynonymous ($N.d_N$) SNPs by the codeml free-ratio model.

Sample	$N.d_N$	$S.d_S$
Chicken [1]	5.9	5.5
Red JF [1]	2.0	4.4
Grey JF [1]	1.0	0
Ceylon JF [1]	3.0	0
Green JF [1]	32.2	16.8
Bamboo partridge	20.0	12.9
Grey francolin	22.7	8.7

[1] On the branch ancestral to the *Gallus* birds, 19.8 nonsynonymous and 6.8 synonymous mutations were observed.

Branch-specific models of evolution implemented with PAML (Yang 1997) were used to investigate evidence of selection among the sequenced lineages. Using the free-ratio model, the branch leading to the *Gallus* genus was determined to have a high ω value (0.92; Figure 5.5), though this cannot be taken as strict evidence of positive selection. Consequently, site-specific models were implemented to investigate whether particular codon sites contributed to the evidence of selection. Model M8, one of the most conservative models of site-specific evolution was determined to be significantly more favoured in comparison to the neutral M7 model ($p = 5 \times 10^{-23}$; Table 5.5). BEB was used to estimate the proportion of sites under positive selection: 48 (9.8%) of the sites had $\omega > 9.5$ – ω values much greater than that expected under neutrality (*ω = 1*; Yang et al. 2005). Under M8, 28 sites were identified to have a BEB posterior probability of at least 95% for *$\omega > 1$* (Table 5.6). There were substitutions between the chicken and red JF sample or genome sequence at six of these sites (5, 517, 547, 590, 628 and 665). PMut predicted four substitutions at these sites would have a neutral effect on protein structure (Table 5.7).

Figure 5.5. Codeml neighbour-joining phylogeny of IL4RA.



Branch lengths were estimated by maximum likelihood under the free-ratio model, which assumed an independent ω-ratio for each branch: these values are displayed. The branch length displayed is 0.1 of the total branch lengths for the tree. The ω for chicken was 0.4181 when sample FJ542675 was used instead of FJ542575. The ω values for grey and Ceylon JF are high because no synonymous SNPs were observed.

Table 5.5. Generated PAML parameters for free-ratio (M1) and significant site-specific test (M2a, M1a; M7, M8) results for IL4RA.

Model	Parameters	Likelihood	$\omega = d_N/d_S$	*2ΔML*	P value
M1	ω = estimated independently for all	-2907.434	See Figure 5.5	-	-
M2a	$\omega_0 = 0$ (90.25%)	-2798.740	$\omega_2 = 10.302$ (9.75%)	102.74	4.88×10^{-23}
M1a	$\omega_0 = 0$ (80.13%)	-2850.114	$\omega_1 = 1$ (19.87%)		
M8 [1]	$\omega_{0\text{-}9} < 0.08$ (9.03% each)	-2798.741	$\omega_{10} = 10.304$ (9.75%)	102.74	4.88×10^{-23}
M7	$\omega_{0\text{-}7} = 0$ (10.0% each)	-2850.115	$\omega_{8,9} = 1$ (10.0%)		

[1] BEB analysis suggests 28 sites where P(ω > 1) > 95.0%. *2ΔML* is twice the difference of the variable model likelihood minus that of the neutral model. The number of degrees of freedom was 2 for these site-specific model LRTs.

Table 5.6. Sites potentially under selection according to BEB analysis of PAML M8 results for the most frequent haplotype.

Base position	P[1]	aA	ω	Se[2]	$P(\omega > 1)$	Exon	Bases	SNP alleles and amino acids[3]
4429-31	3	T	9.983	0.878	0.998	1	ACA	V, F: CCA (P); B: GCA (A)
4435-37	5	F	10.002	0.772	1.000	1	TTT	CK, C: CTT (L); R, V, F: TTC (F); B TTG (F)
4477-79	19	L	9.996	0.807	0.999	1	CTG	V, F: CGC (R); B: CCA (P)
7453-56	23	V	9.951	1.029	0.995	2	GTT	V, F: TTT (F); B: CTT (L)
7534-36	50	E	10.003	0.770	1.000	2	GAA	V, F: CCA (P); B: CGA (R)
7582-84	66	L	10.000	0.786	1.000	2	CTT	V, F: TTT (F); B: AAT (N)
7594-5, 8394	70	R	9.984	0.871	0.998	2, 3	AGA	V: TCA (S); F: ATA (M)
9583-85	125	T	9.999	0.788	1.000	4	ACT	C, B: GCT (A); V, F: TCT (S)
9631-33	141	L	9.771	1.636	0.976	4	TTG	C, G, B: CTG (L); V, F: ATG (M)
9646-48	146	S	9.995	0.811	0.999	4	AGC	V, F: CGC (R); B: GGC (G)
9715-17	169	Q	9.972	0.933	0.997	4	CAA	V, F: CGC (R); B: CCC (P)
9721-23	171	E	9.970	0.942	0.997	4	GAA	V, F: GCA (A); B: GGA (G)
12367-69	418	M	9.661	1.904	0.964	9	ATG	V: CTG (L); B: GTG (V); F: TTG (L)
12628-30	509	A	9.895	1.253	0.989	9	GCA	V, F: GTA (V)
12631-33	510	R	9.966	0.963	0.996	9	AGA	V: AGT (S); B: AGG (R); F: AGC (S)
12652-54	517	H	9.995	0.811	0.999	9	CAC	CK, R: CAT (H); RJF, F, V: CAA (Q); B: AAC (N)
12661-63[5]	520	P	9.110	2.540	0.930	9	CCT	CK, R, RJF: CTT (L)
12742-44[4]	547	I	9.619	2.097	0.954	9	ATA	R, C: TTA (L)
12823-25	574	H	9.941	1.070	0.994	9	CAT	V, F: CAC (H); B, F: CAT (H)
12844-46	581	V	9.976	0.914	0.997	9	GTG	V, F: ATG (M); B: CTG (L)
12871-73	590	G	9.580	2.076	0.956	9	GGC	CK, RJF, B, F, V: AGC (S)
12955-57[4]	618	E	9.622	2.092	0.955	9	GAG	V, F, B: GCG (A)
12979-81	626	S	9.579	2.078	0.956	9	AGC	V: CGC (R); F: GGC (G)
12985-87	628	E	9.934	1.101	0.993	9	GAA	CK: GAG (E); RJF, R, G, C: GAC (D)
13042-44	647	A	9.661	1.902	0.964	9	GCC	V, B, F: GTC (V)
13078-80	659	N	9.727	1.746	0.971	9	AAT	V, F: AAA (K); B: AAC (N)
13096-98	665	R	9.981	0.889	0.998	9	CGA	CK: CAA (Q), TGA (stop); R, G, C: AGA (R): RJF, F: ATA (M); V: ACA (T); B: AAA (K)
13123-25	674	S	9.950	1.033	0.994	9	TCT	V: TGT (C); B: TTT (F); F: TAT (Y)
13138-40	679	A	9.994	0.820	0.999	9	GCA	V: GGC (G); B: GTG (V); F: GGT (G)

[1] Amino acid position. [2] Standard error for ω. [3] CK stands for chicken, B for bamboo partridge, F for grey francolin, V for green JF, C for Ceylon JF, R for red JF, G for grey JF and RJF for the genome sequence. [4] Significant in analysis with divergent sample FJ542675 only. [5] Almost significant.

5.3.3 SNP and population diversity:

Of the 100 SNPs observed among the chicken populations, seven were singletons. In protein-coding regions 17 SNPs were observed: 10 were nonsynonymous and seven were synonymous. Assuming red JF was the primary ancestral origin of diversity at this gene, some replacement mutations between red JF and chicken are potentially associated with the domestication process. Seven nonsynonymous substitutions were segregating at a frequency of 0.55 or more in chicken: F5L, L520P, S590G, L594R, M665R, S670Y and T692S (Table 5.8, Figure 5.6).

Table 5.7. Protein function impacts predicted by PMut for candidate M8 BEB sites under selection that varied between the chicken and the red JF genome sequence.

Base	Amino Acid			Prediction	Score	Certainty	Outcome
Position	Position	Red JF	DA [1]				
4435-37	5	F	L	neutral	0.316	3 [6]	n/s
12652-54	517	Q	H [4]	neutral	0.095	8	neutral
12742-44	547 [2]	L	I	neutral	0.026	9	neutral
12871-73	590	S	G	neutral	0.329	3 [6]	n/s
12985-87	628 [3]	D	E	neutral	0.035	9	neutral
13096-98	665	M	R	neutral	0.495	0 [6]	n/s
13096-98	665	M	Q [5]	neutral	0.119	7	neutral
13096-98	665 [2]	R	Q [5]	neutral	0.510	5 [6]	n/s

[1] Derived allele. [2] Polymorphic between the chicken and the red JF sample. [3] The red JF allele was the same for the genome sequence and sample. [4] The red JF sample and some chickens shared a synonymous SNP at this site. [5] The chicken minor allele at the site. [6] Substitutions where the PMut certainty values ≤ 6 did not have statistical support for the predicted change (n/s).

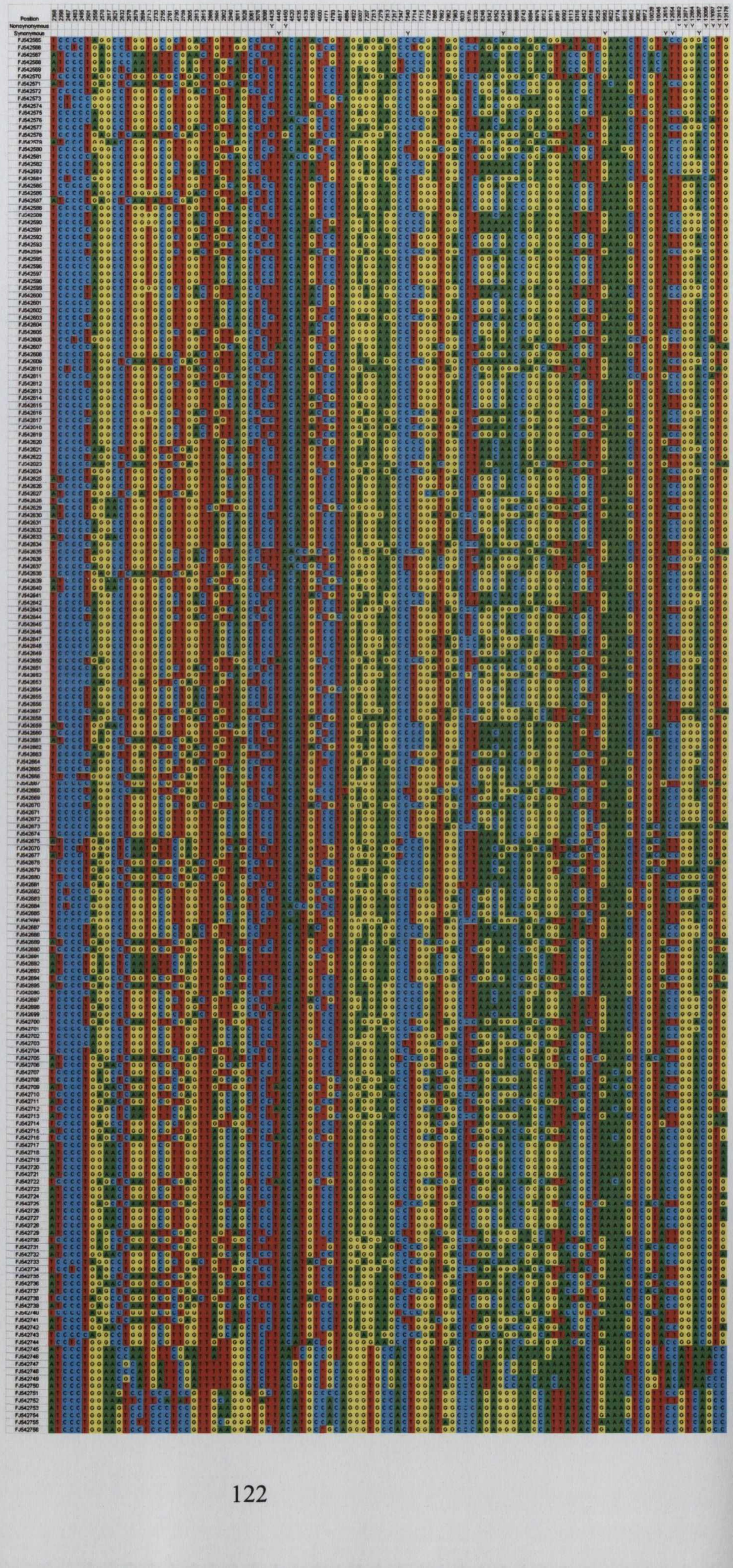
Table 5.8. Frequencies and PMut predicted functional impacts of chicken-red JF genome nonsynonymous SNPs on the IL4RA protein product.

Base	Amino Acid			Prediction	Score	Certainty	N [2]	Outcome
Positions	Position	Red JF	DA [1]					
4435-37	5 [3]	F	L	neutral	0.316	3 [5]	111	n/s
4450-52	10 [3]	T	A	neutral	0.104	7	1	neutral
9622-24	138	N	H	neutral	0.320	3 [5]	1	n/s
12661-63	520	L	P	neutral	0.413	1 [5]	102	n/s
12871-73	590	S	G	neutral	0.329	3 [5]	173	n/s
12883-85	594	L	R	neutral	0.270	4 [5]	174	n/s
13096-98	665	M	R	neutral	0.495	0 [5]	172	n/s
13096-98	665	M	Q	neutral	0.119	7	7	neutral
13096-98	665 [4]	R	Q	neutral	0.510	5 [5]	7	n/s
13096-98	665 [4]	R	stop	-	-	-	1	deleterious
13111-13	670	S	Y	neutral	0.036	9	163	neutral
13111-13	670	S	F	neutral	0.061	8	17	neutral
13111-13	670 [4]	Y	F	neutral	0.023	9	17	neutral
13177-79	692	T	S	neutral	0.037	9	157	neutral
13177-79	692	T	N	neutral	0.060	8	23	neutral
13177-79	692 [4]	S	N	neutral	0.028	9	23	neutral

[1] Derived allele(s) – in some cases this is the most frequent in the chicken samples. [2] N is the number of observed samples with the AA. [3] Amino acid sites 5 and 10 are polymorphic in the outgroup samples as well. [4] Polymorphic within chicken samples only. [5] Substitutions where the PMut certainty values ≤ 6 did not have statistical support for the predicted change.

Figure 5.6. Genotypes at SNP sites polymorphic in the chicken for all samples.

The coding sites are marked as “Y” if nonsynonymous. Samples are from Pakistan (FJ542565-FJ542584), Burkina Faso (FJ542585-FJ542604), Senegal (FJ542605-FJ542624), Sri Lanka (FJ542625-FJ542644), Botswana (FJ542645-FJ542664), Bangladesh (FJ542665-FJ542684), Kenya (FJ542685-FJ542704), Broilers (FJ542705-FJ542744), bamboo partridge (FJ542745-6), grey francolin (FJ542747-8), green JF (FJ542749-50), grey JF (FJ542751-2), Ceylon JF (FJ542753-4) and red JF (FJ542755-6). Bases with nucleotide A are in green, C in blue, G in yellow and T in red.

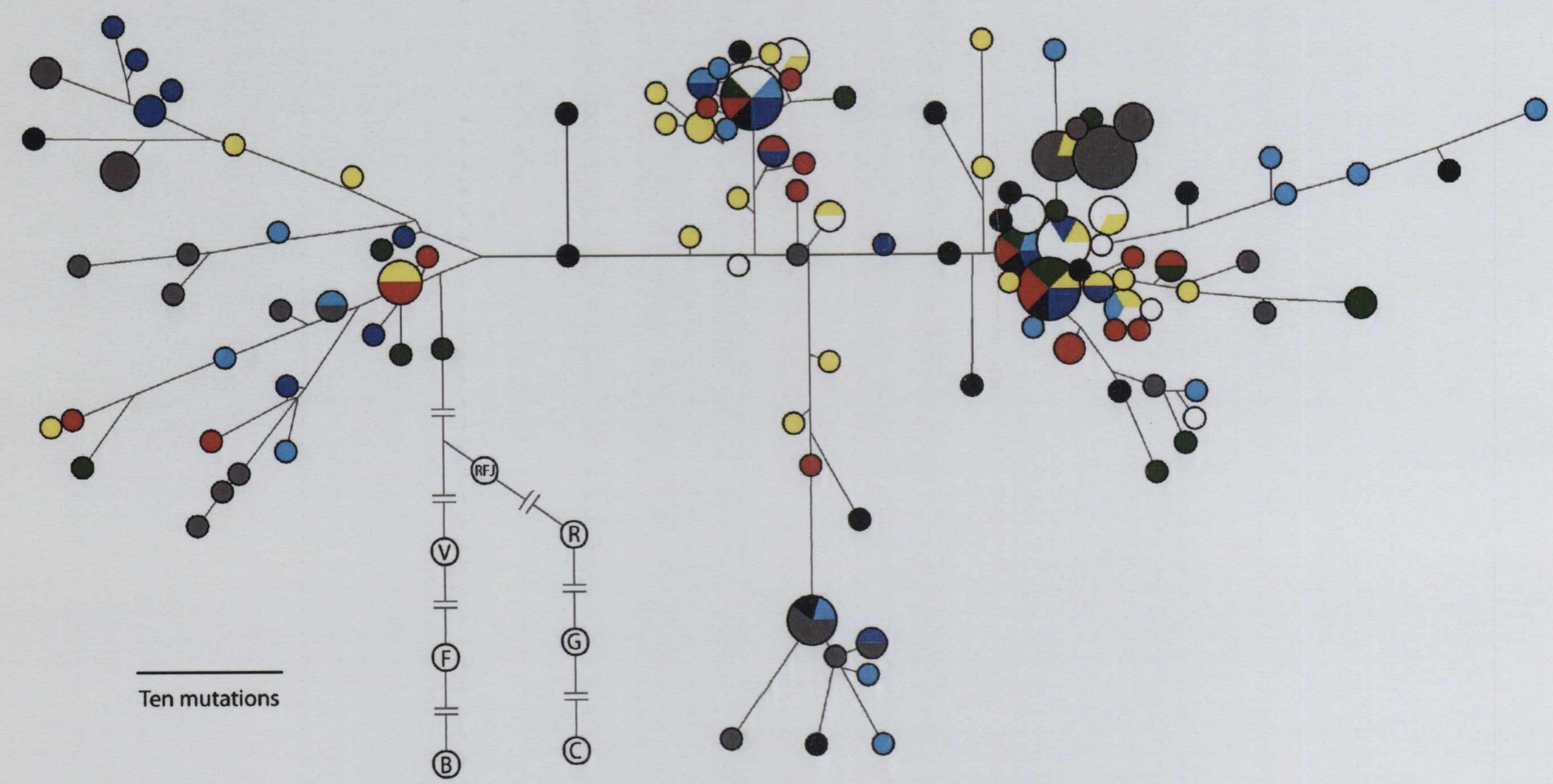

Position
Nonsynonymous
Synonymous

The generation of median-joining networks (Figure 5.7) illustrated a high degree of allele diversity among samples and little geographical structuring among populations. The number of genetically divergent high-frequency haplotypes showed a trend of balanced diversity.

When only the nonsynonymous SNPs were examined, an interesting pattern of dominant haplotypes emerged (Figure 5.8). This picture was obscured when all silent SNPs were included by recombination that dispersed these groups (Figure 5.9). Four haplotypes containing 81% of the 180 genotypes were characterised by substitutions at two sites: F5L and L520P. The 4 alleles possible at these 2 sites (F-L, F-P, L-L and L-P) were present in all 8 populations. No single variant was dominant among the sample genotypes: 32 were F-L, 38 were F-P, 46 were L-L and 64 were L-P. Both sites 5 and 520 showed evidence for positive selection in the site-specific test in codeml (Table 5.6, Figure 5.9). Here, red JF and chicken both shared L520 and P520 alleles as well as F5, but L5 was unique to chicken (Figure 5.10).

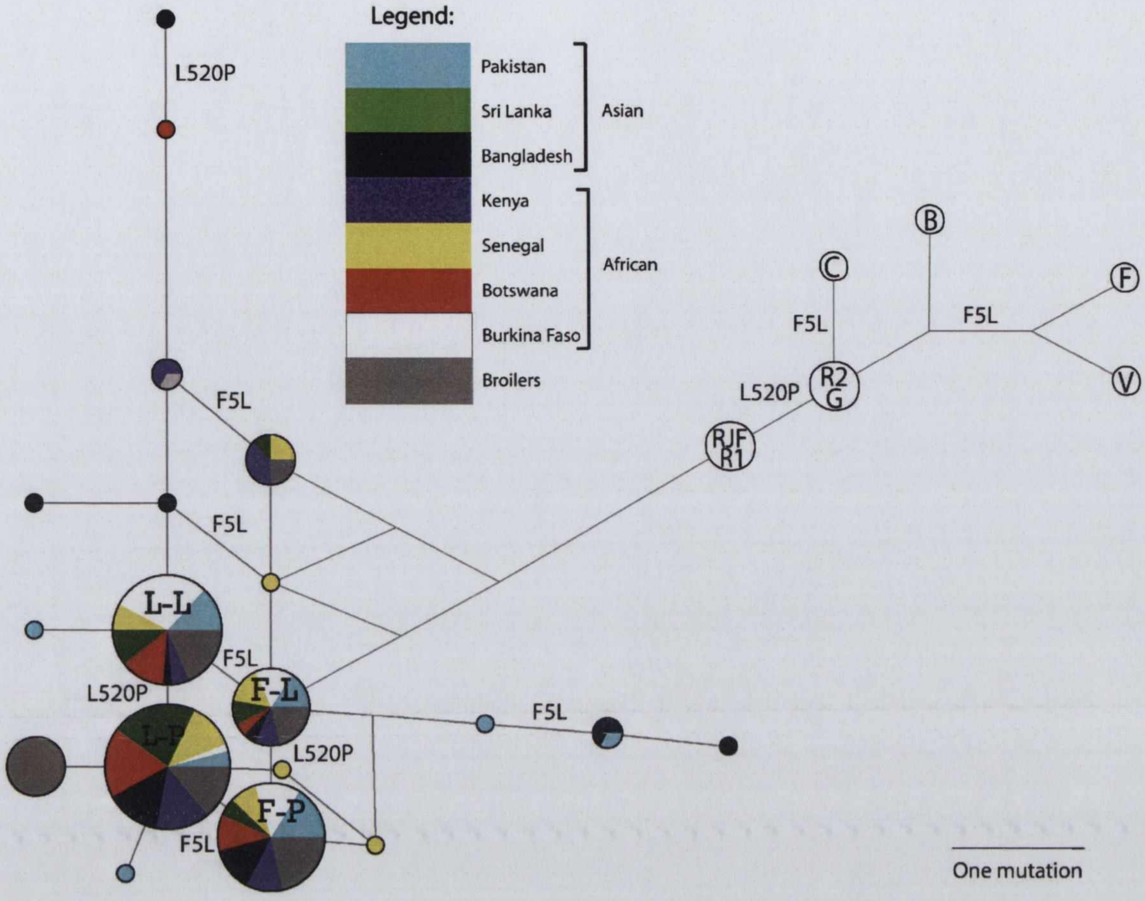
The feature of high population diversity and little geographic partitioning in the networks was apparent in the analysis of variation using AMOVA with the Arlequin package (Schneider et al. 2000). This assessed the extent of partitioning of diversity at different levels of population structure. Most variation lay within the populations (94.1%, $p < 1 \times 10^{-5}$), a trend observed in other studies of chicken populations (Kanginakudru et al. 2008); the remainder partitioned between the populations (1.8%, p = 0.060) and the continents (4.1%, p = 0.033).

Figure 5.7. Median-joining network of chicken haplotypes for all SNPs.



Populations are denoted in the legend. Branch lengths are proportional to the number of mutational differences between haplotypes. The outgroup samples are represented by the colourless nodes. Outgroup sample branch lengths were considerably reduced in order to show the details of the chicken population network. V represents the green JF sequences; F the grey francolin; B the bamboo partridge; G the grey JF; C the Ceylon JF; R the red JF; and RJF the red JF genome sequence.

Figure 5.8. Median-joining network of chicken haplotypes for nonsynonymous SNPs.

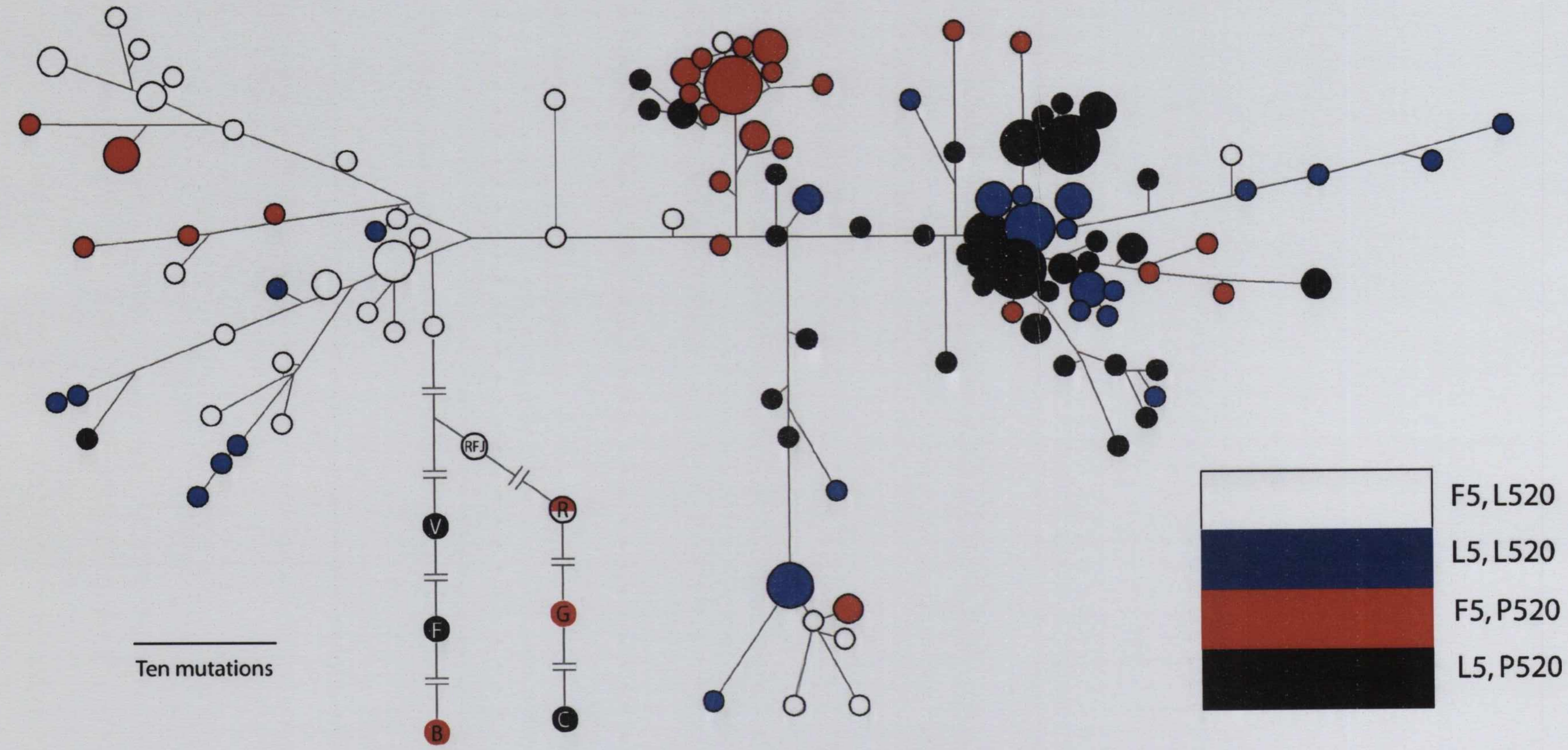


Populations are denoted in the legend. Branch lengths are proportional to the number of mutational differences between haplotypes. The outgroup samples are represented by the colourless nodes. V represents the green JF sequences; F the grey francolin; B the bamboo partridge; G the grey JF; C the Ceylon JF; R1 and R2 the red JF; and RJF the genome sequence.

5.3.4 Summary statistics and tests of neutrality:

There was further evidence for the trend of elevated allelic diversity: 115 haplotypes were observed in just 180 genotypes, which was reflected in the high *Hd* value (Table 5.9). The significantly positive Tajima's *D* in Asia and Africa (Table 5.9) and in each of their populations (Table 5.10) was paralleled by a highly negative Fay and Wu's *H*, an indicator of an excess of derived alleles. Together, these metrics suggested a clear tendency for alleles to rise to mid- or high- frequency levels. Tests on the protein-coding portion of the gene alone indicated a significantly negative Fay and Wu's *H* (-3.02, $p < 0.05$; Table 5.9) and a less positive Tajima's D (0.61); the latter may be a consequence of stronger conservation in coding regions, which appeared to limit diversity, except at sites 5 and 520.

Figure 5.9. Median-joining networks of haplotypes for all SNPs classed according to substitutions at key amino acids (F5L, L520P) from Figure 5.8.



The four possible genotypes at these positions are denoted in the legend. Branch lengths are proportional to the number of mutational differences between haplotypes. Outgroup sample branch lengths were considerably reduced in order to show the details of the chicken population network. V represents the green JF sequences; F the grey francolin; B the bamboo partridge; G the grey JF; C the Ceylon JF; R the red JF sample genotypes; and RJF the genome sequence.

Table 5.9. SNP data, summary statistics and tests of neutrality.

All sites	N[1]	S[2]	Hap[3]	*Hd*[4]	π[5]	θ_w[6]	Tajima's *D*	Fu & Li's *D*	Fu & Li's *F*	Fay & Wu's *H*	Fu's F_s
All 90	90	100	115	0.990	5.19	3.37	1.69	1.22	1.86	-21.40	-34.06
P value			*<0.001*	*<0.001*	*n/s*	*0.001*	*<0.001*	*0.027*	*<0.001*	*0.010*	*0.015*
Asia	30	95	51	0.993	5.37	3.89	1.32	1.25	1.68	-19.45	-16.06
P value			*<0.001*	*0.007*	*n/s*	*0.005*	*<0.001*	*<0.001*	*0.009*	*0.008*	*0.012*
Africa	40	86	53	0.983	4.72	3.36	1.36	1.11	1.55	-27.34	-10.13
P value			*n/s*	*n/s*	*n/s*	*0.002*	*0.001*	*0.031*	*0.002*	*0.002*	*n/s*
Broilers	20	79	23	0.944	5.14	3.51	1.69	2.15	2.47	-10.85	1.38
P value			*0.010*	*0.003*	*n/s*	*0.018*	*<0.001*	*<0.001*	*<0.001*	*n/s*	*0.007*

[1] Number of chickens sampled. [2] SNPs. [3] Haplotypes. [4] Haplotype diversity. [5] Mean pairwise differences per kb. [6] Watterson's estimator per kb. Only p values generated by simulations < 0.05 are given; p > 0.05 are denoted "n/s". Of the 5,298 sites resequenced in total, 1,472 were coding and 3,380 were noncoding sites
Fu's F_S was highly negative, signifying an excess of rare alleles. Nucleotide, haplotype and SNP diversity were all higher in Asia than in Africa as expected, despite sampling fewer birds in Asia (30) than in Africa (40).

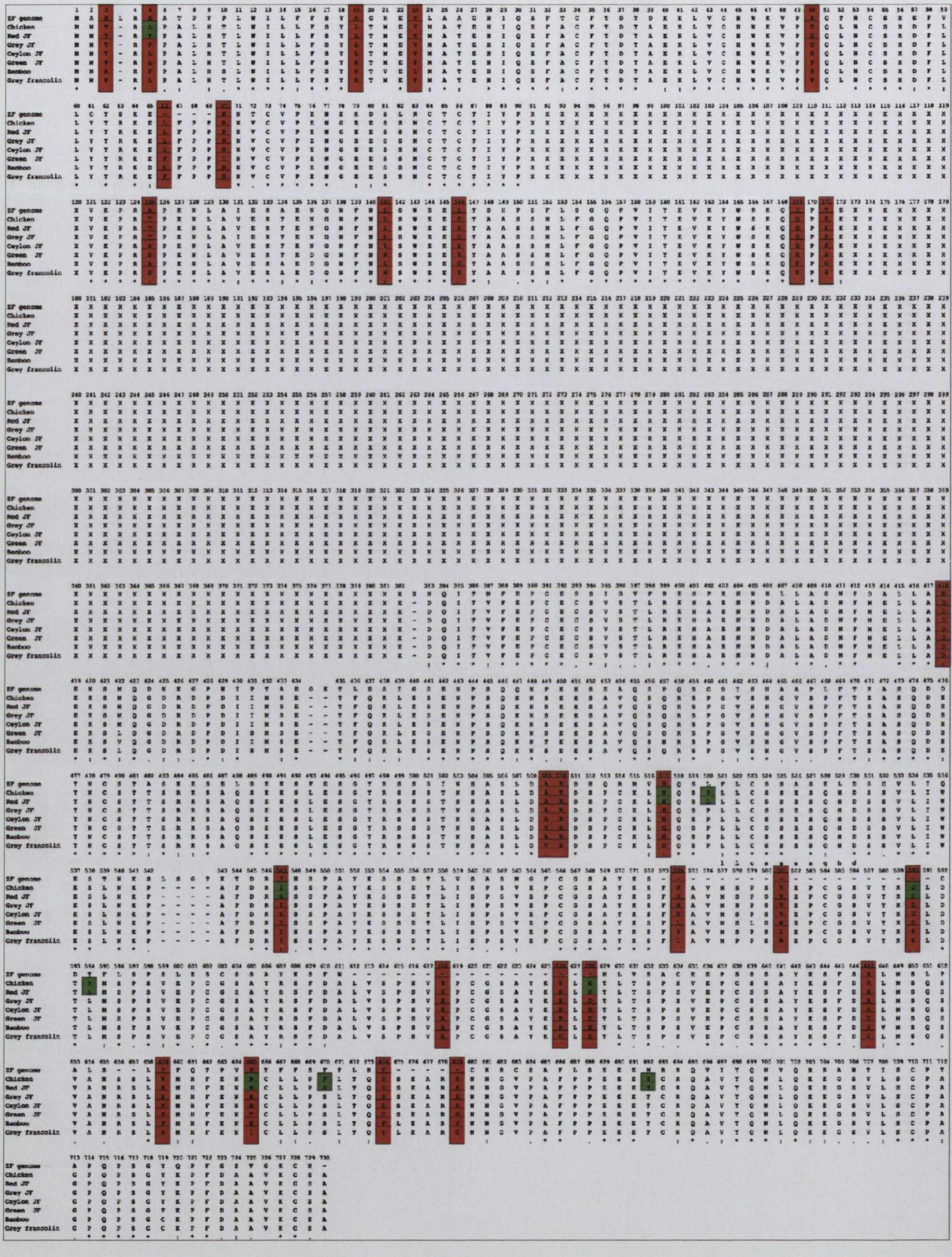
Table 5.10. Tajima's *D* and Fay & Wu's *H* for each Asian and African population.

Continent	Asia			Africa			
Population	Bangladesh	Pakistan	Sri Lanka	Botswana	Burkina Faso	Senegal	Kenya
SNPs	82	86	73	74	51	70	71
Tajima's *D*	0.85	1.10	0.86	0.90	1.21	1.03	1.81
P value	*0.032*	*0.006*	*0.033*	*0.015*	*0.004*	*0.012*	*<0.001*
Fay & Wu's *H*	-14.33	-17.58	-22.23	-22.32	-21.85	-14.22	-12.41
P value	*0.031*	*0.027*	*0.004*	*0.003*	*<0.001*	*0.016*	*0.037*

For each population, 10 chickens were sampled.

Moderate recombination was detected at IL4RA: for the calculated value of the recombination rate (*R*) coalescent simulations showed the minimum number of recombination events (R_M) was significantly high among all groups (Table 5.11). The effects of recombination were apparent in the disruption of the phylogenetic network groups (Figure 5.8, Figure 5.10, Figure 5.11).

Figure 5.10. A multiple sequence alignment of zebra finch and other bird samples IL4RA protein-coding sequences.



Legend to Figure 5.10: Sites marked were candidates for selection according to PAML M8 BEB results (red), and had differences in the chicken populations compared to the red JF genome or samples (green). Regions marked with X were not resequenced. Bamboo refers to the bamboo partridge. Chicken had 2 alleles (F, L) at site 5; red JF, grey JF and bamboo partridge all had F; and Ceylon JF, green JF and grey francolin had L. At site 520 the alleles segregating in chicken (L, P) were present in chicken and red JF, and though zebra finch genome had L, the remaining birds all had P.

Table 5.11. Recombination at IL4RA according the percentage GC content, Hudson's R and R_M and Kelly's Z_{nS} per kb from DnaSP.

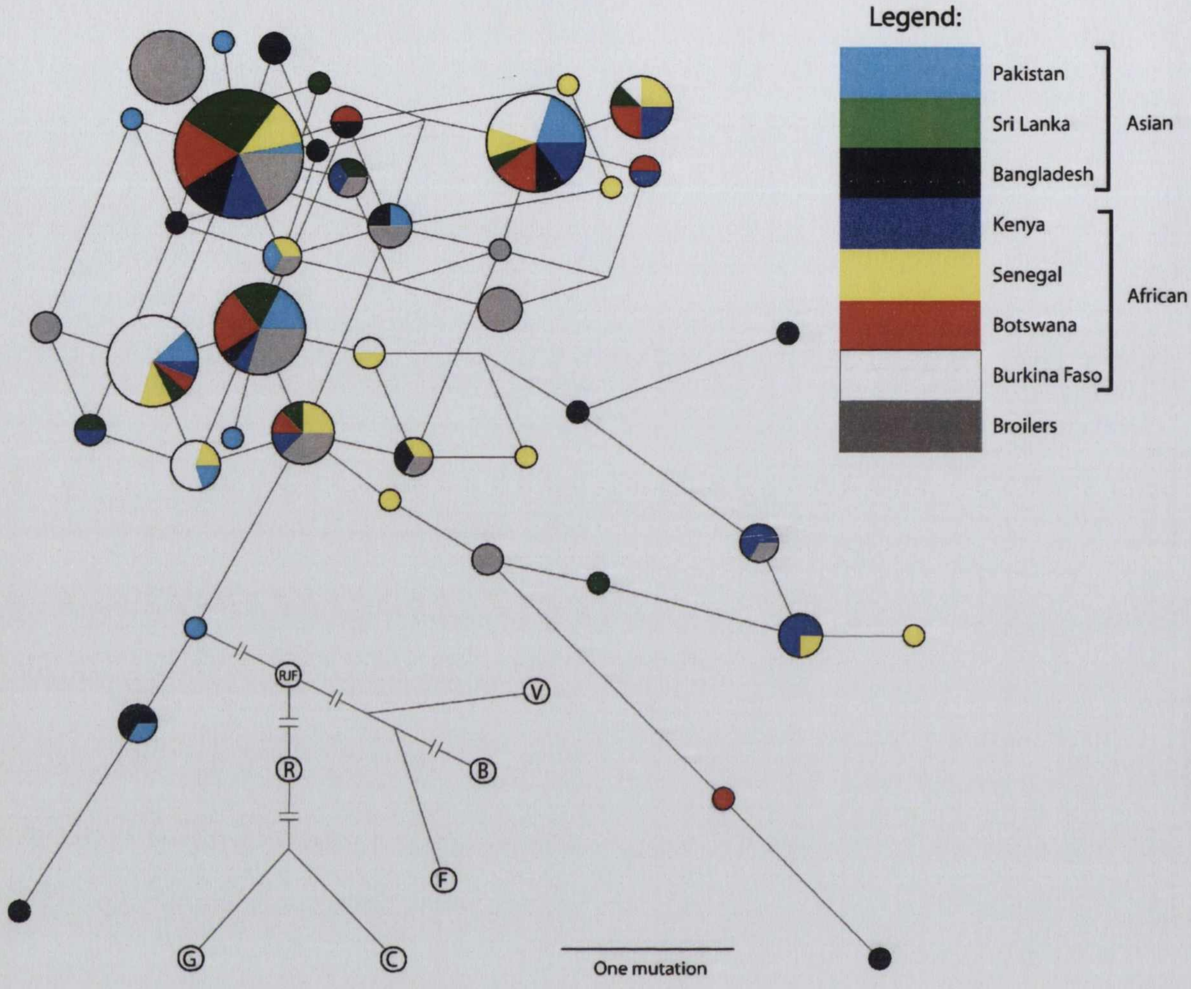
GC content (%)				R_M [1]				Z_{nS} [2]
Total	Coding	Non-coding	R	All	Asia	Africa	Broilers	
44.5	46.3	43.8	33.60	35	27	21	17	66.13

[1] Coalescent simulations in DnaSP: $p < 0.001$ for all, $p < 0.001$ for Asia, $p = 0.004$ for Africa and $p = 0.013$ for broilers. [2] Coalescent $p = 0.017$.

Evidence of non-neutral evolution was evident from the McDonald-Kreitman test results. The McDonald-Kreitman test examines the relative ratios of fixed and non-fixed nonsynonymous differences to fixed and non-fixed silent changes between species (D_N/D_L versus P_N/P_L; McDonald & Kreitman 1991). Background selection may explain a rate of fixation of nonsynonymous differences much lower than that for silent substitutions. Alternatively, if there is a significant excess of fixation of nonsynonymous changes compared to silent ones, then directional selection may be present.

The chicken genotypes were tested against the red JF genome sequence and also against the outgroup samples. Both tests showed a dearth of nonsynonymous substitutions fixed between species ($p < 0.01$ with the genome sequence, $p = 0.04$ with all six outgroups, $p < 0.05$ with all outgroups except red JF; Table 5.12), indicating that purifying selection affected the evolution of this gene. For chicken versus the genome sequence $D_N/D_L = 0.108$ and $P_N/P_L = 1.000$, and for chicken versus the outgroup sequences $D_N/D_L = 0.159$ and $P_N/P_L = 0.667$. Both these observations indicated more extensive conservation at nonsynonymous sites on the lineages separating chicken from the genome sequence, and separating chicken from the JF, bamboo partridge and grey francolin.

Figure 5.11. Median-joining network of chicken haplotypes for coding SNPs.



Populations are denoted in the legend. Branch lengths are proportional to the number of mutational differences between haplotypes. The outgroup samples are represented by the colourless nodes. Most outgroup sample branch lengths were considerably reduced in order to show the details of the chicken population network. V represents the green JF sequences; F the grey francolin; B the bamboo partridge; G the grey JF; C the Ceylon JF; R the red JF; and RJF the red JF genome sequence.

While an excess of nonsynonymous substitutions within chicken may be a sign of adaptive evolution (Eyre-Walker 2002), it can also be a sign of relaxed purifying selective constraint. McDonald-Kreitman tests for positive selection would need to be conducted in a more robust manner, which would entail testing relative ratios for fixed and polymorphic substitutions at nonsynonymous versus synonymous rather than silent sites. When such tests were implemented for the samples in Table 5.12, there was no significant evidence of selection. Given that an excess of silent compared to nonsynonymous substitutions on the ancestral JF lineage was detected, it is likely that the McDonald-Kreitman tests were significant due to conservation at nonsynonymous sites.

Table 5.12. McDonald-Kreitman tests between the chicken populations and the red JF genome sequence and the outgroup samples.

Samples tested	Substitution Type	Intraspecific	Interspecies	P value
Chicken vs genome sequence	Silent	6	93	0.002
	Nonsynonymous	6	10	
Chicken vs 6 outgroups	Silent	6	447	0.040
	Nonsynonymous	4	71	
Chicken vs 5 [1] outgroups	Silent	6	436	0.045
	Nonsynonymous	4	72	

[1] Not including the red JF sequences.

5.3.5 Zebra finch IL4RA gene on chr14:

The GenBank zebra finch IL4RA mRNAs used in this analysis included 5' UTR and perhaps leader sequence, like the chicken copy. The zebra finch IL4RA coding region starts with a 69 base-long first exon at position chr14:16,260,749. Other identifiable orthologous coding regions to chicken are exon 2 at 16,264,259-402, exon 3 at 16,265,272-427, exon 4 at 16,266,443-604, exon 5 at 16,267,141-299, exon 8 at 16,268,959-9,036 and exon 9 at 16,269,132-71,277. Regions for exons 6, 7 and 10 were not clear as the zebra finch mRNA sequences were short and the divergence between chicken and zebra finch was high at the 3' end of the gene: this was reflected in the number of segregating polymorphism in the chicken samples.

5.4 Discussion

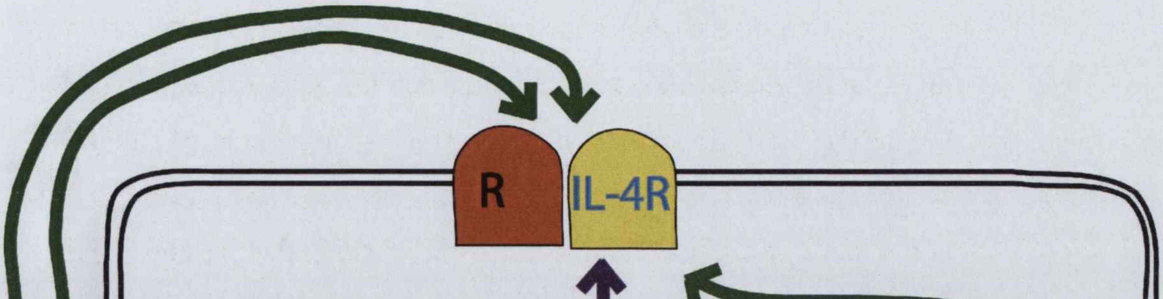
5.4.1 Identifying IL4RA as a candidate for resequencing:

A pairwise comparison of $\omega = d_N/d_S$ in chicken and zebra finch genes identified 12 genes, including IL4RA, as having an elevated rate of nonsynonymous substitutions, suggesting they were possibly subject to positive selection (Yang & Nielsen 2002). It is possible some of these genes were candidates due to relaxed selective constraint, which has been observed in other domestic species (Cruz et al. 2008), however, the low general pattern of ω values suggested most genes were conserved.

Interestingly, the two other chicken immune genes identified with this pairwise comparison method (Pibf and Pias2) have human orthologs that interact with IL4RA and its signalling pathway (Figure 5.12; KEGG www.genome.jp – pathway 04060). Human Pibf is an immunoregulatory factor expressed during embryo development that regulates T_H1 and T_H2 cytokine production balance by binding the IL4RA and an anchored Pibf receptor chain, which activates Jak1 to phosphorylate STAT6 (Anderle et al. 2008). Normally, activated STAT6 proteins dimerise and translocate to the nucleus, where they activate T_H2 cytokines (Liu et al. 1998, Kozma et al. 2006). However, human Pias proteins may prevent cytokine activation by inhibiting STAT proteins in the nucleus (Chung et al. 1997, Shuai & Liu 2005). Thus, the 3 immune genes identified by this method not only are expected to interact in the same pathway but also are likely to have crucial roles in modulating the immune response of chickens to viruses, bacteria and parasites.

Due to its important role in the host immune response and evidence of selection in humans, IL4RA was resequenced in six closely related birds and subsequently in 70 global village chickens and 20 commercial broilers. An analysis of sequence data from these six related species identified a large number of sites likely to be subject to positive selection, supporting the initial detection of IL4RA as a candidate gene undergoing adaptive evolution. Probable confounding factors in these results, however, are the complex domestication history of these populations and high rate of recombination identified at this locus.

Figure 5.12. Simplified cellular schematic of the interactions between human IL4RA, Pibf and Pias2 gene products.



The three candidate genes products are in blue; other proteins and cell components are black. Genes are pink and the black arrows indicate their expression. Green arrows indicate activation by binding; the red arrow represents Pias2 inhibiting activated STAT. Human IL4RA (yellow) forms part of a transmembrane receptor (purple arrow) with other interleukin receptor chains IL13RA1, IL2RG or the Pibf receptor ("R", orange).

The identification of chicken IL4RA is of particular interest given the vital role played by its human ortholog as a regulator of IgE production and T_H2 cell differentiation (Wu et al. 2001, Liu et al. 2004). In mammals, the γ_c chain dimerises with the IL13RA1 or the IL4RA before binding IL4 and IL13, suggesting that the chicken IL4RA can interact with IL13 as well (Junttila et al. 2008). The critical role of human IL4RA in the immune response is evidenced by its differential expression during particular infections and the association of its variability with disease susceptibility; it facilitates gastrointestinal nematode clearance (Horsnell 2007) and its expression is upregulated in response to HIV-1 infection (Puri et al. 1992). Variation in human IL4RA has been shown to affect signal transduction (Kruse et al. 1999) and to modulate T_H1/T_H2 balance in the blood (Youn et al. 2000), as well as contributing to various allergies (Shirakawa et al. 2000) and to mumps virus infection susceptibility (Dhiman et al. 2008). Selection at IL4RA in human populations may be driven by different T_H1 (viral and bacterial) and T_H2 (parasitical) immune responses to pathogens (Wu et al. 2001), and the dysregulation of such components of immunity may be associated with atopy (Hershey et al. 1997).

5.4.2 The origin of diversity at IL4RA:

Although nucleotide diversity at this gene (5.19 per kb) was comparable to that observed between red JF and domestic fowl (5.36 per kb on average; International Chicken Genome Sequencing Consortium 2004), the substantial excess of haplotypes was suggestive of non-neutral evolution. Despite this, the significantly positive Fu and Li's *D* and *F* values show that there was a relative deficit of singletons (Fu & Li 1993). A deficit of rare alleles in commercial chicken lines has been observed in other studies comparing wild and standard breeds (Muir et al. 2008). In this study, the *Hd* and Fu's F_S values highlighted this rare allele deficiency in the commercial broilers, in contrast with the excess of haplotypes in the Asian and African samples. In addition, the significantly high R_M values indicated that some recombinant alleles were present in the populations, implying either relaxed selective constraint or adaptive processes favouring allelic diversity.

Tajima's *D* compares the proportions of low- to medium-frequency alleles and is an indicator of directional or purifying selection when negative, and balancing selection when positive (Tajima 1989). Fay and Wu's *H* measures the relative frequency of derived alleles, which increases when there are more high-frequency haplotypes (Fay & Wu 2000). The observed surplus of mid- and high-frequency haplotypes at the IL4RA locus has generated highly significant *D* and *H* values that are more extreme than those observed by other studies of disease-associated chicken genes (Berlin et al. 2008) – however, *D* and *H* are likely to be affected by demographic aspects of chicken history and the pooling of samples (Carlson et al. 2005).

The networks were diffused into several divergent high-frequency haplotype clusters with high intra-population diversity. A distinctive set of balanced alleles was apparent when silent substitutions were removed. The signal of balanced diversity in the chicken populations appeared to centre around two nonsynonymous substitutions: F5L and L520P. All four variants at these two sites were segregating in the 8 populations surveyed at similar frequencies. Site-specific models of evolution identified both these sites as likely subject to selection across species.

An alignment of the chicken and human IL4RA protein sequences identified the amino acid positions orthologous to sites 5 and 520 in chicken (Figure 5.2). The site orthologous to 520 is segregating in humans (C431R, rs1805012; Deichmann et al. 1997, Lozano et al. 2001) at an intermediate frequency of over 10% in the population (Landi et al. 2007), similar to the chicken polymorphism here. Substitution C431R is in the cytoplasmic domain of the receptor and is linked with better survival from gliomas in humans (Wrensch et al. 2006). The human amino acid position orthologous to chicken site 5 is conserved (F10) and is located in the signal peptide of the protein, indicating that the L5 chicken variant might affect expression activation of the receptor protein.

There is a series of shared population genetic properties between chicken and human IL4RA that may be the result of equivalent functional roles for each. The genes possess comparable positive McDonald-Kreitman test results and Tajima's *D* values as well as sharing orthologous high-frequency nonsynonymous SNPs (L520P and C431R). And given that several amino acid substitutions in IL4RA affect disease susceptibility in humans (see Franjkovic et al. 2005) the variability at nonsynonymous substitution sites in chickens is likely to be of biological importance.

The balanced and elevated variation and possible selective processes at chicken IL4RA may be in response to common pathogens and the range of pleiotropic roles that the receptor plays in facilitating cytokine binding in the innate immune response. The trend of high diversity fuelled by balancing selection has been seen at other chicken immune genes including MHC-B (Worley et al. 2005), Mx (Seyama et al. 2006, Berlin et al. 2008) and IL1B (Downing et al. 2009a), which initially suggests that immune system genes may maintain high diversity in order to respond to a wide array of pathogens.

Another explanation for the observed elevated balanced diversity is that multiple domestications of red JF and genetic introgressions of other JF have both enhanced and distorted variation at this locus. The lack of observed geographic structure, which has also been observed at other chicken genes may be in part a consequence of this. There are likely to have been multiple events of chicken domestication in south and south-east Asia (Liu et al. 2006, Oka et al. 2007, Fumihito et al. 1996). And though the red JF is the main source of chicken genetic diversity (International Chicken Genome Sequencing Consortium 2004, Fumihito et al. 1994), genetic introgressions have come from other wild JF (Eriksson et al. 2008, Silva et al. 2008, Nishibori et al. 2005). Wild red JF and domestic village strains are closely related (Yang & Nielsen 2002, Berthouly et al. 2009), indicating that introgressions of red JF may have continued after domestication. Here, networks of IL4RA indicated that red JF is the most closely related wild relative to the domestic chicken. This does not exclude the possibility of multiple contributions of different genetic sources of JF. If admixture of different sources occurred sufficiently early through trading and migration (Berthouly et al. 2009, Muchadeyi et al. 2008, West & Zhou 1989) this may explain the presence of the four alleles at the two nonsynonymous sites in each population. Regardless of whether this signal of high and balanced diversity is from biological pleiotropy or from multiple origins, it is persisting, indicating that it may have an important role in current chicken immunity.

5.5 Conclusion

This study shows evidence for high and balanced diversity at the chicken IL4RA gene, which was initially identified through the evaluation of the rate of nonsynonymous to synonymous substitutions in pairwise comparisons of chicken and zebra finch orthologs. This strategy incorporated functional and literature information to detect a suitable gene for resequencing in African, Asian and commercial chicken samples, as well as in related JF and bird species. Haplotype networks, tests of neutrality and summary statistics indicated a signal of balanced nonsynonymous polymorphisms at two sites in the IL4RA gene. Networks showed that red JF is the primary source of diversity at this gene. The elevated and balanced diversity present in all the populations might be a result of the chicken's history of multiple domestications, introgressions (2008) and subsequent admixture of different types.

However, the identification of two potentially functionally significant SNPs as fulcrums of the balancing signal suggest that the functions of IL4RA in the immune system may affected by selective processes for specific allelic variants in response to new pathogenic challenges during domestication.

Publication

This chapter formed the basis for a publication in BMC Evolutionary Biology 9(1):136 in 2009 entitled “Bioinformatic discovery and population-level validation of selection at the chicken interleukin-4 receptor alpha-chain gene”. The authors are: Downing T, Lynn DJ, Connell S, Lloyd AT, Bhuiyan AKFH, Silva P, Naqvi A, Sanfo R, Sow RS, Podisi B, O’Farrelly C, Hanotte O, Bradley DG.

CHAPTER 6

Variation in chicken populations may affect the enzymatic activity of lysozyme

6.1 Introduction

The innate component of the immune system forms the initial response to any pathogen invasion and also acts to stimulate the adaptive immune system, which takes additional time to respond (Medzhitov & Janeway 2000). This is the most ancient part of immunity and is present in plants as well as animals – by contrast, only vertebrates have an adaptive side (Janeway & Medzhitov 2002). Consequently, innate immune defences sustain selective pressure from novel pathogen adaptations to develop effective and swift mechanisms to fight disease. Improving the innate immune response of chickens to disease can enhance the adaptive component as well, and has direct relevance to ongoing research in commercial broilers (Swaggerty et al. 2009).

6.1.1 Chicken lysozyme as a model gene:

Hydrolytic enzymes that can disrupt key parasitic, bacterial or viral cell components are an important part of innate chicken defence systems. Lysozyme is one such protein whose bactericidal activity was initially identified serendipitously in human nasal mucous (Fleming 1922). It operates by hydrolysing peptidoglycan and chitodextrin, both of which are components of gram positive cell membranes (Holler et al. 1975a, Holler et al. 1975b), and is also effective against gram negative bacteria (Pellegrini et al. 1997). Lysozyme's unique combination of properties, fast crystallisation and high concentration in egg-white, from which it can be purified easily, made it a model protein for primary investigations of spatial structure using X-ray crystallography (Blake et al. 1965, Johnson & Phillips 1965). The catalytic mechanism of lysozyme hydrolysis was determined using this technique (Strynadka & James 1991). The gene also served as a model for exploring gene regulation in complex organisms (Bonifer et al. 1997).

6.1.2 Chicken lysozyme's role in disease resistance:

Lysozyme's elevated expression *in ovo* highlights the importance of its role, at which stage of development innate immune mechanisms are vital because the adaptive immune system is not yet fully developed (Sippel *et al.* 1978, Ask *et al.* 2007). The study of chicken lysozyme has unveiled insights into human susceptibility to disease: for example, the resistance of certain *Streptococcus* cell walls to chicken lysozyme and to human leukocyte enzymes is determined by the same set of compositional factors (Glick et al. 1972). Chicken lysozyme can help resist the infections of *Micrococcus lysodeikticus*, *Flavobacterium columnare* and *Edwardsiella tarda* when expressed by zebra fish (Yazawa et al. 2006).

Specific sites in lysozyme are responsible for different components of its activity. Amino acids 98 to 112, which are part of a helix-loop-helix (HLH) domain at sites 87 to 114, have antimicrobial activity against *Serratia*, *Micrococcus* and *Staphylococcus* species without having muramidase activity (Pellegrini et al. 1997). This HLH domain is active against gram negative and gram positive bacteria, as well as certain fungi (Ibrahim et al. 2001). Substitutions at other sites can change the catalytic effectiveness of the enzyme by either enhancing (Goto et al. 2008) or diminishing it (Harada et al. 2008, Kawamura et al. 2008). Mutant forms of lysozyme can compromise immune system function in rabbits (Prieur et al. 1974). Additionally, variants of lysozyme are associated with amyloidosis in humans (Pepys et al. 1993). Thus it is possible that chicken lysozyme has sites that have been subject to sharp evolutionary pressures during its evolution.

Here, diversity present at the gene was explored by resequencing it in chicken populations and related species, and found one nonsynonymous substitution segregating at an intermediate frequency. Tests indicated that this site and one other nonsynonymous change fixed between red JF and chicken were spatially close to the key catalytic sites.

6.2 Methods

6.2.1 Sample collection:

The same chicken samples from Asia and Africa from the International Livestock Research Institute (Kenya) in Chapters 4 and 5 were used. Samples from red, grey and Ceylon JF from Wallslough Farm (Ireland); green JF, bamboo partridge and grey francolin from the Californian Academy of Sciences; and 20 commercial broilers from Manor Farms (Ireland) were also surveyed. More details of the samples are listed in Chapters 4 (Asian, African and outgroups) and 5 (broilers). The DNA was isolated from the samples using a phenol-chloroform extraction following a proteinase K digestion.

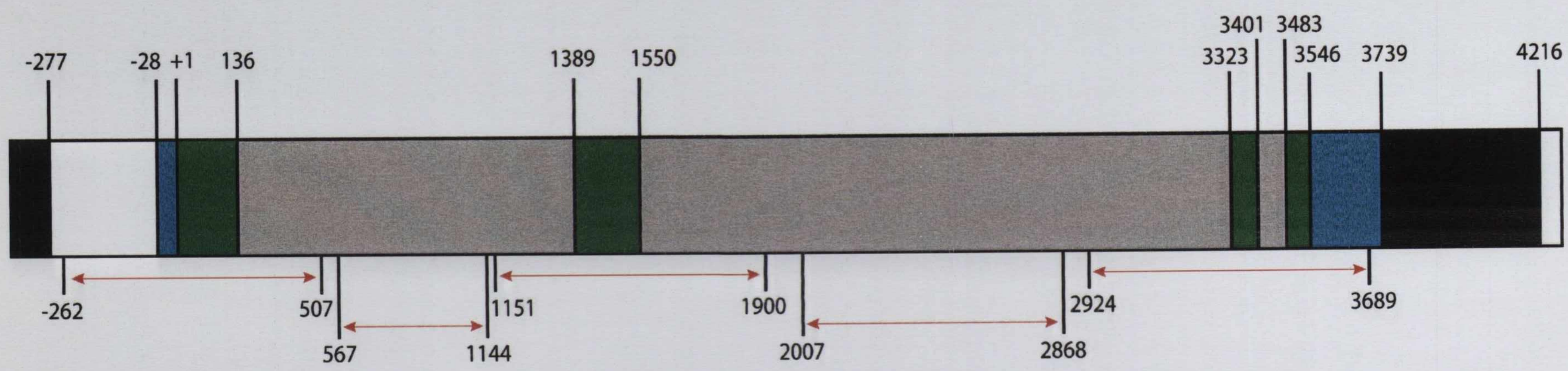
6.2.2 Sequence determination and acquisition:

The UCSC, Ensembl and GenBank Map View (http://www.ncbi.nlm.nih.gov/projects/mapview/) browsers were used to map the gene structure using GenBank contig NW_001471454 and reference assembly NC_006101. PCR primer sequences (Table 6.1) were designed using Primer3 according to the parameters in Chapter 4 and were created by VHBio. Five amplicons covering 3,726 bp of the gene were successfully amplified by PCR (Table 6.2) for the 96 samples (Figure 6.1). The forward and reverse PCR product sequences were determined by Agowa.

Table 6.1. Sets of primer pair sequences and their associated optimal PCR parameters.

Amplicon	Size (bp)	Orientation	T_M (°C)	[MgCl] (mM)	Primer Sequences
1	770	Forward	59	25	TGATGAACAATGGCTATGCAGT
		Reverse			TTCTCCCCCACTACTCCTTGTA
2	578	Forward	54	25	TACAAGGAGTAGTGGGGGAGAA
		Reverse			ATAAATTCCAGCGTGCTTTTGT
3	750	Forward	55	20	ACAAAAGCACGCTGGAATTTAT
		Reverse			CTTCACTAGTGGGATGGGAAAG
4	862	Forward	62	15	TAAGGTGAAACGACACTCATGG
		Reverse			CTACAACCTCTCTGGGCAGTCT
5	766	Forward	58	20	CTATGAGAGTGGTGAGGTGCTG
		Reverse			AAGGCGTTTGCGTATAGTCG

Figure 6.1. Lysozyme gene structure.



Exons are shown in green, introns in grey and amplicon regions by the red arrows. The UTRs are shown in blue and non-sequenced regions are in black. The numbers shown represent the base positions in relation to the GenBank entry for the mRNA sequence.

Table 6.2. PCR cycle program used for each pair of primers.

Step	Temp. (^{o}C)	Duration (min)
1	95	15
2	95	0.5
3	T_M	0.75
4	72	1
5	72	15

T_M is the annealing temperature as listed in Table 6.1. Steps 2 to 4 were repeated 33 times in sequence.

6.2.3 Sequence assembly and haplotype reconstruction:

The DNA base sequencing generated chromatograms that were assembled into contigs using the Phred-Phrap-Consed-Polyphred pipeline programs Phrap v0.990319 and Phred v0.020425.c (Ewing & Green 1998, Ewing et al. 1998). Bases were called, SNPs were selected and sequences were assembled as detailed in Chapter 5.

PHASE version 2.1.1 (Stephens et al. 2001) was used to reconstruct the haplotypes and to infer any missing haplotypes. A list of the genotypes for each sample was collated. Perl scripts were used to remove any sequence sites where there was inadequate coverage across all populations or continents, or sub-standard base quality scores, or insufficient coverage for either forward or reverse sequences. The sequences were exported to Mega (version 4.0.2, Tamura et al. 2007) to convert the data to formats useable by other software packages. Haplotypes were assigned using PHASE and these were cross-referenced with haplotypes generated by Arlequin (Schneider *et al.* 2000) to ensure consistency: the haplotypes generated by both were identical. The genome sequence was used to align the resequenced regions so that relative exon positions could be confirmed by MGalign version 3.1 (Lee et al. 2003).

6.2.4 Data analysis:

AMOVA tests (Excoffier et al. 1992) were conducted on all sites using Arlequin, with 1,000 permutations (Schneider et al. 2000). See Chapter 4 for details. Median-joining haplotype networks were constructed using Network version 4.2.0.1 (Bandelt et al. 1999) software.

DnaSP 4.0 (Rozas & Rozas 1999, Rozas et al. 2003) was used to analyse the polymorphic characteristics of the data and to perform a series of population genetic analyses – see Chapter 4 for complete methods on each metric. The numbers and types of SNPs were assessed. Nucleotide diversity was measured using π (Tajima 1983). The haplotype diversity (*Hd*, Depaulis & Veuille 1998, Equation 2), the number of haplotypes, Z_{nS} (Kelly 1997, Equation 3) and $\theta_W = 4N_e\mu$ (Watterson 1975, Equation 4) were determined. The four gamete test for the minimum number of recombination events (R_M; Hudson & Kaplan 1985, Equation 5) and *R* (the degree of recombination; Hudson 1987, Equation 6) were calculated, as was the GC content.

A set of summary statistics were used to identify departures from neutrality using coalescent simulations: Fu and Li's *D* and *F* (Fu & Li 1993, Equations 8 and 9), Tajima's *D* (Tajima 1989, Equation 7), Fu's F_s (Fu 1993, Equation 11) and Fay and Wu's *H* (Fay & Wu 2000, Equation 10). These were implemented in DnaSP as discussed in Chapter 4. These simulations generated empirical distributions with which the statistical values were compared to determine the extent of their deviation from neutrality. It is an indication of non-neutral evolution if the observed values lie at the extremes of the distribution.

The above statistics evaluate intraspecific variation, so these were combined with an interspecies examination of evidence for selection using the ratio of the relative rate of nonsynonymous mutations (d_N) to the relative rate of synonymous mutations (d_S) in the protein-coding portion of the gene. This was calculated as d_N/d_S (ω) for models using the codeml implementation of PAML 3.15 package (for details see Chapter 4; Yang 1997). The free-ratio model (M1) and tests for positive selection on specific sites (models M2a vs M1a, M8 vs M7) were implemented.

6.2.5 Protein spatial modelling and impact prediction:

The proximity of amino acids of interest can yield information regarding their possible effect on the effectiveness of the enzyme. The spatial relationships of the amino acids that were polymorphic or are involved in catalysis were examined in a three-dimensional model of chicken lysozyme displayed in RasMol 2.7.4.2 (http://www.openrasmol.org/software/rasmol/). Euclidean distances between α-carbon atoms recorded the lysozyme protein database (PDB) file were calculated for each of the sites of interest (PDB ID 3B6L; Michaux et al. 2008).

Predictions to estimate the extent of functional impact of each nonsynonymous substitution were conducted using SIFT (Ng and Henikoff 2003), PMut (Ferrer-Costa et al. 2005) and PolyPhen (Ramensky et al. 2002); Chapter 4 details the software more fully.

Multiple sequence alignments using T-Coffee (Notredame et al. 2000) of human and bird versions of the gene were completed to identify patterns in avian variation. For most species, only the active regions of the protein sequences were available.

6.3 Results

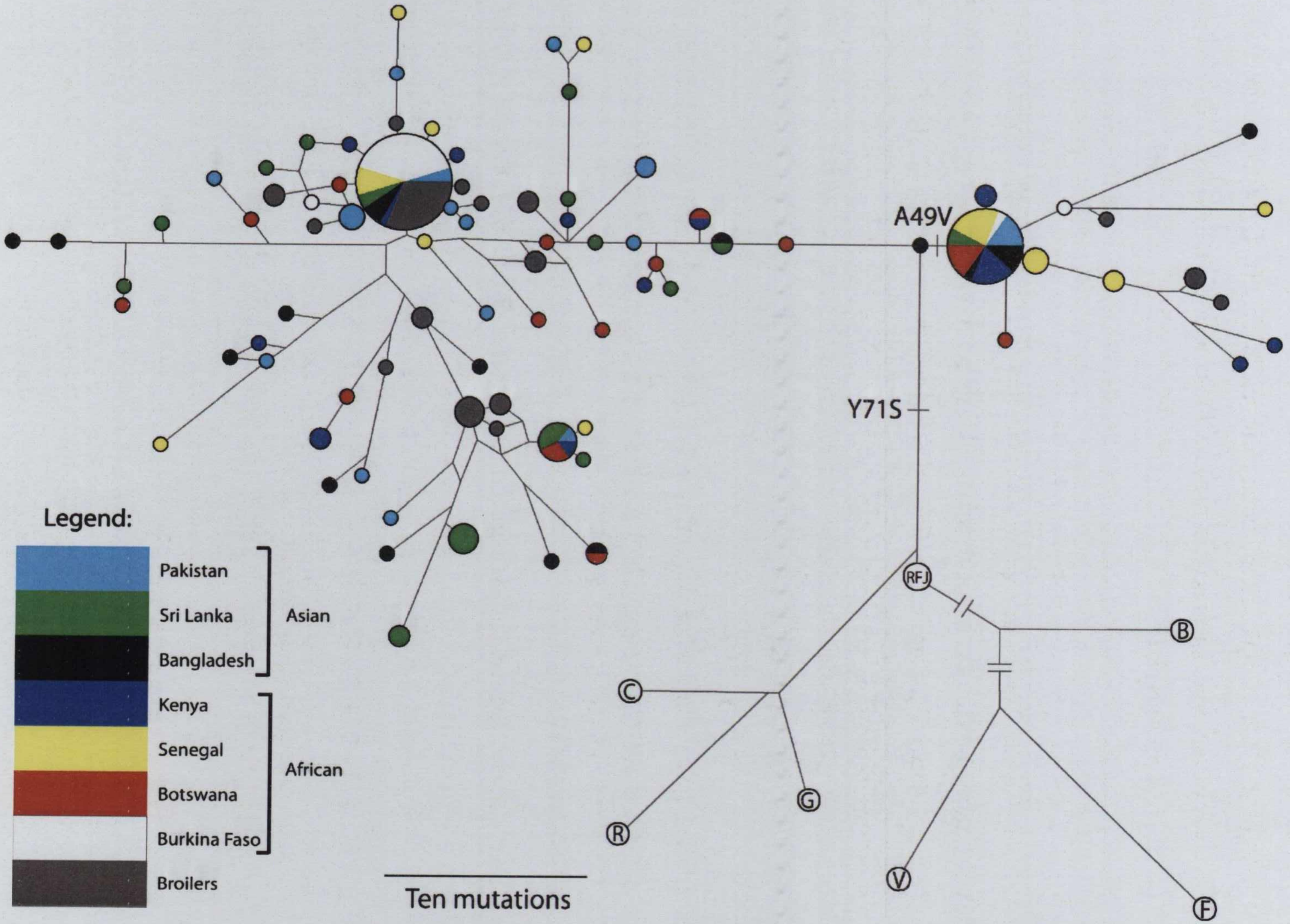
Lysozyme is a crucial innate immune system enzyme expressed at high levels in developing chicken eggs. It was resequenced in seven village chicken populations from Africa (Botswana, Burkina Faso, Kenya and Senegal) and Asia (Bangladesh, Pakistan and Sri Lanka) and a set of broilers – a total of 90 samples for each gene. Additionally, four samples in the same genus as chicken (red, grey, Ceylon and green JF) and two outgroup species closely related to chicken (bamboo partridge and grey francolin) were examined.

6.3.1 Chicken population and jungle fowl diversity:

Of 59 SNPs discovered among domestic chicken genotypes for this gene, only one was a non-singleton coding SNP (cSNP). Slightly more SNPs are found in Asia (54) than in Africa (53), despite sampling more in the latter than the former. Broilers had significantly fewer SNPs (37), in part because only 20 of such samples were analysed. Only three cSNPs were discovered: two of these were singleton alleles, one of which was nonsynonymous (S71F) – the other was synonymous. The one non-singleton cSNP was a nonsynonymous substitution (A49V) at base 1398 in exon 2 and defined the two most numerous haplotypes of 84 observed in total in a median-joining network (Figure 6.2). In this network, the red JF genome sequence was the most proximal JF genotype to the chicken samples.

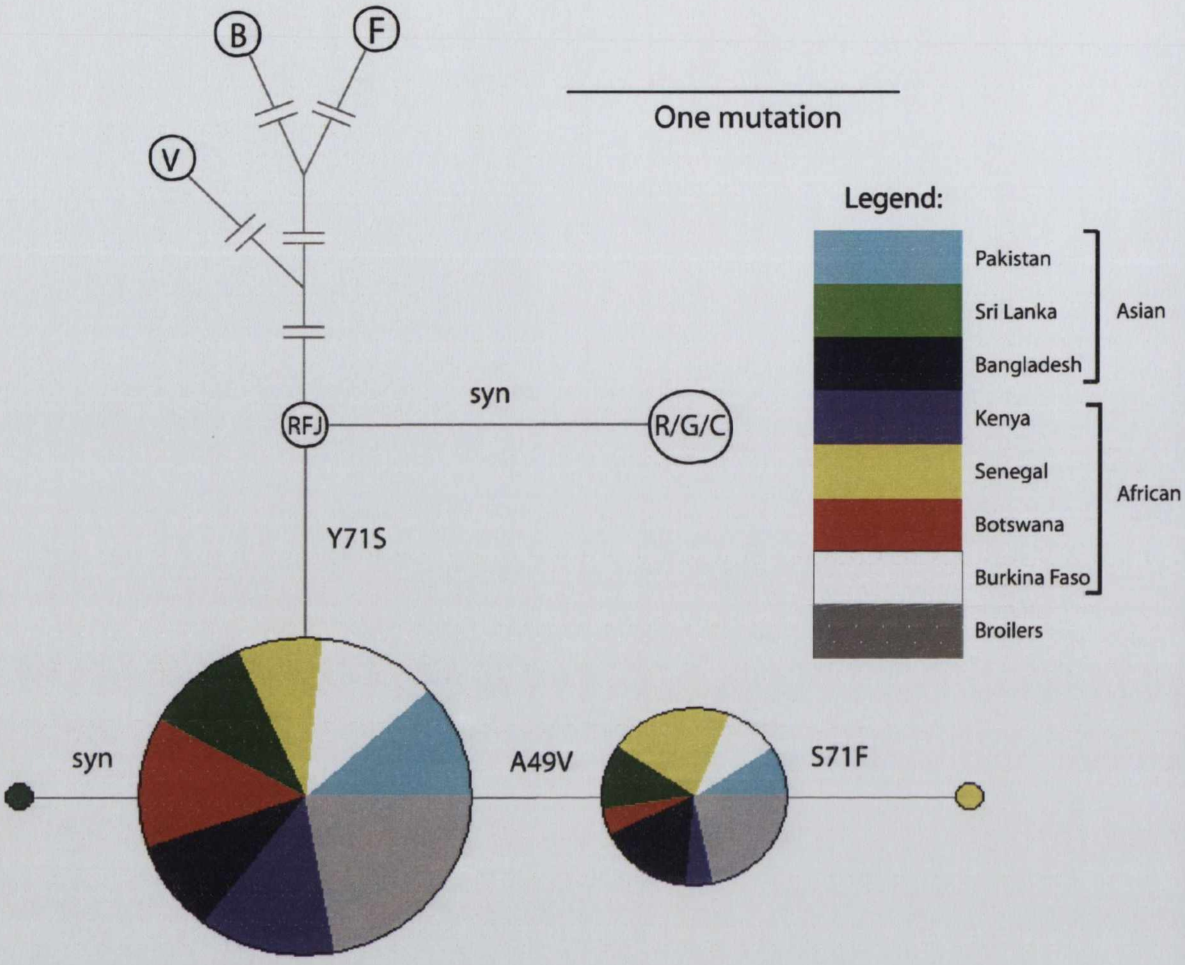
When only cSNPs were used to construct a network (Figure 6.3), this substitution alone separated the two principal alleles that were present in all 8 chicken populations. Substitution Y71S at site 1464 was the solitary cSNP distinguishing the red JF genome sequence and the chicken haplotypes. The grey, red and Ceylon JF were separated from the reference red JF genome sequence by a single synonymous SNP at base 1699. Because of the extensive coding sequence conservation, segregating cSNPs may have more functionally relevant implications.
Analysis of variation at different levels of population structure with Arlequin (Schneider et al. 2000) using AMOVA (Excoffier et al. 1992) showed the high allelic diversity observed in the phylogenetic networks was partitioned within the populations (90.59%, $p < 10^{-5}$), and among populations (9.41%, $p < 10^{-5}$); but not among the continents.

Figure 6.2. Median-joining phylogenetic network of chicken and outgroup haplotypes.



Populations are denoted in the legend. Branch lengths are proportional to the number of mutational differences between haplotypes. The outgroup samples are represented by the colourless nodes: their branch lengths were considerably reduced in order to show the details of the chicken population network. V represents the green JF sequences; F the grey francolin; B the bamboo partridge; G the grey JF; C the Ceylon JF; R the red JF; and RJF the red JF genome sequence.

Figure 6.3. Median-joining network of chicken population haplotypes for coding SNPs only.



Populations are denoted in the legend. Branch lengths are proportional to the number of mutational differences between haplotypes. The outgroup samples are represented by the colourless nodes; their branch lengths were considerably reduced in order to show the details of the chicken population network. V represents the green JF sequences; F the grey francolin; B the bamboo partridge; G the grey JF; C the Ceylon JF; R the red JF; and RJF the genome sequence, which was just one cSNP in distance to the major haplotype. The red, grey and Ceylon JF coding sequences were identical.

6.3.2 Intraspecific patterns of variability:

Significantly high allelic variation was observed at the gene: this was supported by the AMOVA analysis and coalescent simulations incorporating recombination that evaluated the degree of deviation from neutrality of the observed data for a number of statistics, including the haplotype diversity (*Hd* = 0.923; Table 6.3) and Fu's F_S (-34.48). A relative deficit of singletons shown by the positive Fu and Li's *D* (1.42) and *F* (1.34) suggested that such alleles were not the cause of the elevated allele variation.

Table 6.3. Gene data, summary statistics and tests of neutrality.

All sites	S [1]	H [2]	*Hd* [3]	Π [4]	θ_w [5]	Tajima's	Fu & Li's		Fay &	Fu's
						D	*D*	*F*	Wu's *H*	F_s
All	59	84	0.923	3.56	2.96	0.618	1.42	1.34	-14.08	-34.48
P value		*<0.001*	*0.006*	*ns*	*ns*	*ns*	*0.022*	*0.016*	*0.002*	*0.008*
Asia	54	39	0.969	4.12	3.31	0.824	0.99	1.14	-10.69	-9.93
P value		*ns*	*ns*	*ns*	*0.026*	*ns*	*ns*	*0.036*	*0.028*	*ns*
Africa	53	36	0.886	3.35	2.96	0.440	1.05	1.06	-13.92	-5.48
P value		*ns*	*0.001*	*ns*	*ns*	*ns*	*ns*	*0.045*	*0.001*	*ns*
Broilers	37	17	0.881	2.81	2.42	0.566	1.27	1.20	-16.43	0.15
P value		*0.032*	*0.013*	*ns*	*ns*	*ns*	*0.028*	*0.044*	*0.001*	*0.011*

30 chickens from Asia, 40 from African and 20 broilers were sampled. All sites (3,726 bp) includes 440 bp of coding and 3286 bp of noncoding sequence. [1] Number of SNPs. [2] Number of haplotypes. [3] Haplotype diversity. [4] Mean number of pairwise differences per kb between sequences. [5] Watterson's estimator per kb. P values are generated by 1,000 DnaSP coalescent simulations for given recombination rate; only those whose $p < 0.05$ are given.

A positive Tajima's *D* (0.618) and negative Fay and Wu's *H* (-14.08) supported a trend of the elevated variation balanced around the two most numerous haplotypes in the network diagrams (Figure 6.2), which appeared to centre on substitution A49V. Although a significantly high minimum number of recombination events ($R_M = 25$) suggested that some new haplotypes created by recombination were preserved (Table 6.4, Table 6.5), these recombinants were not maintained at nonsynonymous sites (Figure 6.3).

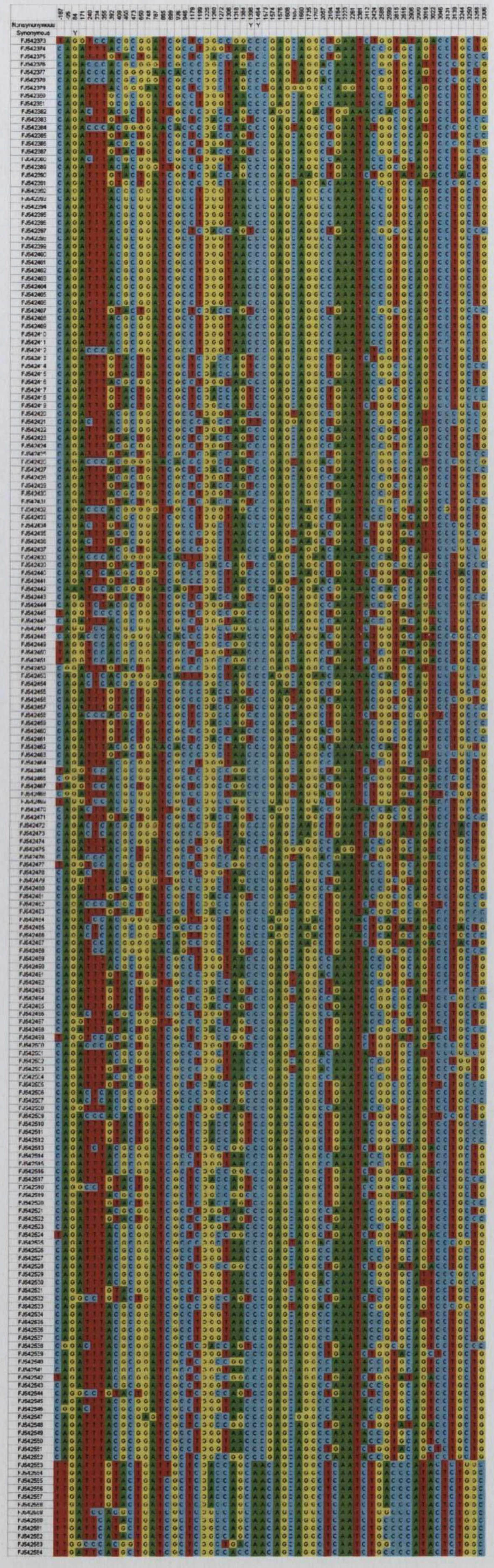
Table 6.4. Recombination according the percentage GC content, Hudson's *R* and R_M and Kelly's $Z_n s$ per kb from DnaSP.

GC content (%)			*R*	R_M [1]				Z_{nS} [2]
Total	Coding	Non-coding		All	Asia	Africa	Broilers	
49.0	57.4	47.8	24.199	25	19	15	8	6.528

[1] Coalescent $p < 0.001$ for all values except Europe ($p = 0.039$). [2] Coalescent p value not significant.

Table 6.5. Genotypes at SNP sites polymorphic in chicken samples.

The coding sites are marked as "Y" if coding and the leftmost column denotes if these are synonymous or nonsynonymous. Bases with nucleotide A are in green, C in blue, G in yellow and T in red. Samples are from Pakistan (FJ542373-FJ542392), Burkina Faso (FJ542393-FJ542412), Senegal (FJ542413-FJ542432), Sri Lanka (FJ542433-FJ542452), Botswana (FJ542453-FJ542472), Bangladesh (FJ542473-FJ542492), Kenya (FJ542493-FJ542512), Broilers (FJ542513-FJ542552), bamboo partridge (FJ542553-4), grey francolin (FJ542555-6), green JF (FJ542557-8), grey JF (FJ542559-60), Ceylon JF (FJ542561-2) and red JF (FJ542563-4).



6.3.3 Interspecies tests for selection:

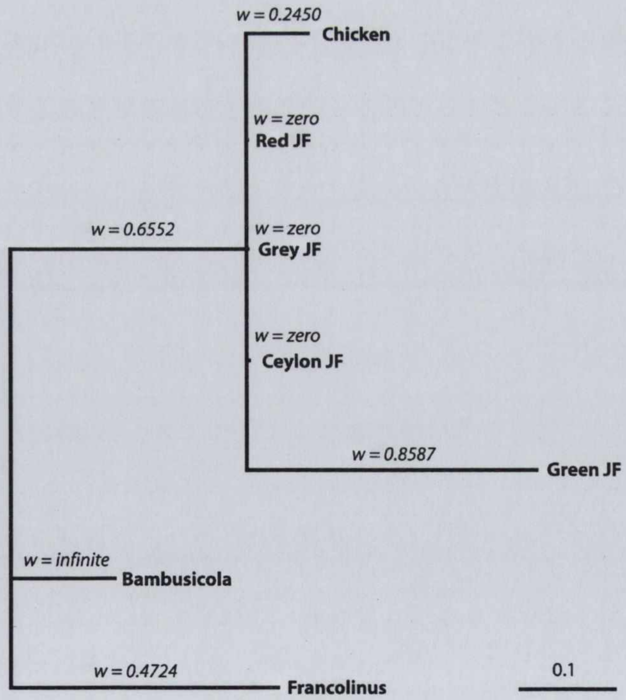
Using the codeml implementation of PAML 3.15 package, d_N/d_S (ω) was calculated for the free-ratio model, which allows independently estimated ω values based on probabilities of the genealogical positions of mutations (Table 6.6) and branch lengths for each lineage (Figure 6.4; Yang 1997).

Table 6.6. Estimated distribution of synonymous ($S.d_S$) and nonsynonymous ($N.d_N$) SNPs by the codeml free-ratio model.

Sample	$N.d_N$	$S.d_S$
Chicken [1]	1.0	1.0
Red JF [1]	0	0
Grey JF [1]	0	0
Ceylon JF [1]	0	0
Green JF [1]	7.2	2.1
Bambusicola	3.4	0
Francolinus	5.7	3.0

[1] 5.5 nonsynonymous and 2.1 synonymous changes are in the branch ancestral to the *Gallus* genus.

Figure 6.4. Neighbour-joining phylogeny constructed using codeml.



Branch lengths are estimated by maximum likelihood under the codeml free-ratio model, which assumed an independent ω-ratio for each branch: these are displayed above each branch. The branch length displayed is 0.1 of the total branch lengths for the tree.

Site specific models estimate ω for each site across the whole coding sequence for a neutral model (M7, $0 \geq \omega \leq 1$) and a variable model (M8) that allows for $\omega > 1$ as well as $0 \geq \omega \leq 1$. (Yang 1997). Likelihood ratio tests (LRT) were conducted with 2 degrees of freedom for M2a vs M1a and M7 vs M8 for the seven resequenced species. LRTs performed between these model pairs showed M8 was significantly more likely than M7 according to a χ^2 distribution, as model M2a over M1a (both $p = 5 \times 10^{-4}$, Table 6.7). Using a random sites BEB model, a significantly high posterior probability of $\omega > 1$ ($p > 0.95$) indicated positive selection at candidate sites 57, 70, 72 and 96 (Table 6.8; Nielsen and Yang 1998, Yang et al. 2005).

Table 6.7. Generated PAML parameters used and output for significant test results for the major coding allele.

Model	Parameters	Likelihood	$\omega = d_N/d_S$	$2\Delta ML$	P value
M1	ω = estimated independently for all	-723.6036	As per Figure 6.4	-	-
M2a	$\omega_0 = 0$ (87.46%)	-747.9517	$\omega_2 = 7.065$ (12.55%), $\omega_1 = 1$ (0%)	15.025	0.0005
M1a	$\omega_0 = 0$ (73.68%)	-755.4643	$\omega_1 = 1$ (26.32%)		
M8	$\omega_{0\text{-}9} = 0$ (8.76% each)	-747.9517	$\omega_{10} = 7.07$ (12.55%)	15.191	0.0005
M7	$\omega_{0\text{-}6} = 0$ (10.00% each)	-755.5473	$\omega_{7\text{-}9} = 1$ (10.00%)		

2ΔML is twice the difference between the likelihoods of the variable and the neutral models.

Table 6.8. Sites potentially under positive selection according to BEB analysis of PAML M8 results.

Sites	ω	ω SE	P(ω>1)	Major allele	Outgroup	Derived allele	Bases
57	7.358	2.398	0.954	AAC (Asparagine)	B: F, V:	AAG (Lysine) AAA (Lysine)	1421-23
70	7.635	2.038	0.992	GAC (Aspartate)	B: F, V:	TAC (Tyrosine) CAC (Histidine)	1460-62
72	7.407	2.347	0.960	GGA (Glycine)	B, F, V:	GAA (Glutamate)	1466-68
96	7.634	2.039	0.992	ATC (Isoleucine)	B, V: F:	TTC (Phenylalanine) GTC (Valine)	1538-40

SE is standard error. Outgrp is the outgroup sample species. All sites are in exon 2.

Given the dearth of coding sequence variability among the resequenced species, a diverse set of lysozyme protein sequences for a range of species: human, zebra finch, turkey (*Meleagris gallopavo*) and birds from the *Phasianidae* family (chicken; copper (*Syrmaticus soemmerringii)*, kalij (*Lophura leucomelanos)*, and golden pheasant

(*Chrysolophus pictus)*; bobwhite (*Colinus virginianus*) and Japanese quail (*Coturnix japonica*)) were aligned using T-Coffee (Notredame et al. 2000). This alignment showed that all *Phasianidae* had a different amino acid (A) at site 49 compared to the zebra finch and human (L; Figure 6.5). L49 in human is shared by mammals *Papio anubis* (NP_001106112), *Pan troglodytes* (NP_001009073), *Macaca mulatta* (XP_001117369), *Mus musculus* (NP_038618) and *Bos taurus* (NP_001071297). Alignments of lysozyme at a genetic lysozyme resource webpage (http://lysozyme.co.uk/) indicate that many species sequenced for the gene have either A49 or L49. Site 71 was conserved in all samples (Y), except the domestic chicken (S), indicating that while site 49 has evolved in the avian lineage, site 71 appeared to be altered in chicken alone.

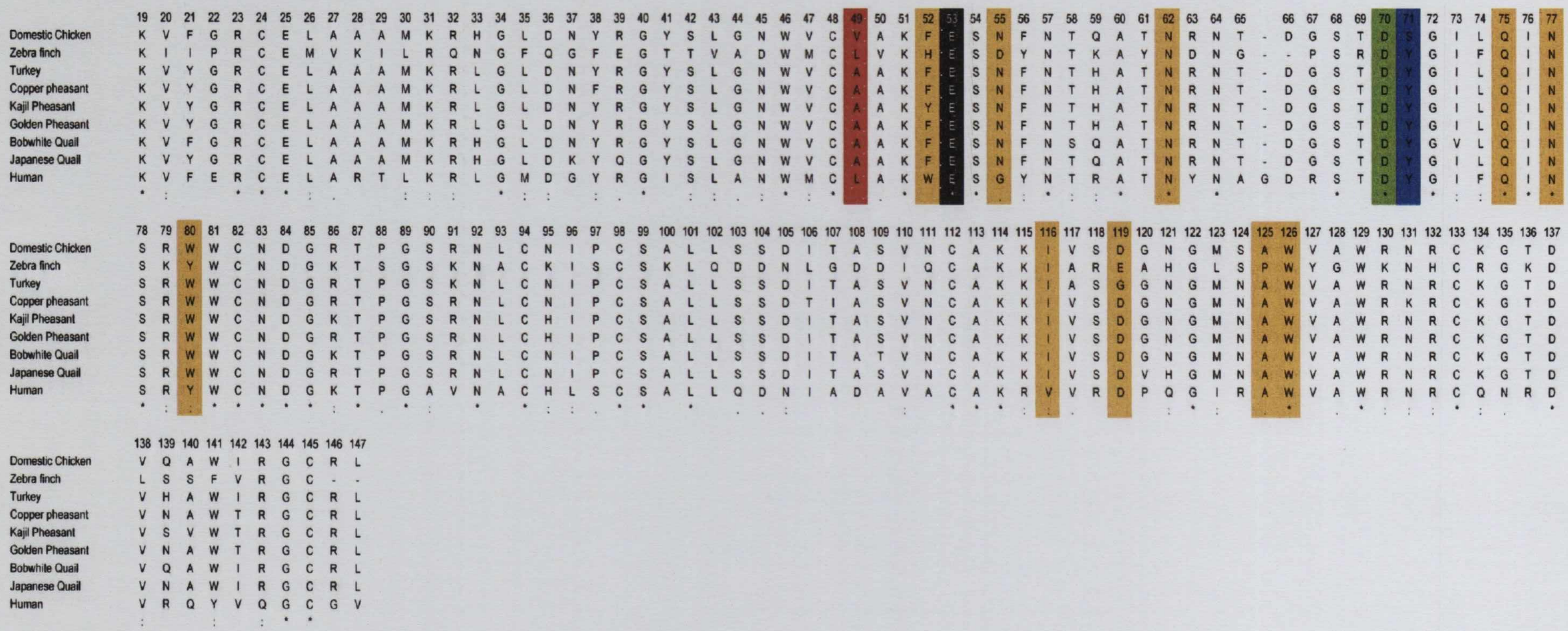
Predictions to estimate the extent of functional impact for each nonsynonymous substitution with PMut, SIFT and Polyphen were generally not effective or had conflicting results for certain sites, most likely due to the high protein sequence divergence between chicken and the species with which it was compared. The prediction outcomes were classed as not determined, probably neutral or probably deleterious, depending on the output of all three programs (Table 6.9).

Table 6.9. Predicted functional impacts of different nonsynonymous SNPs on the protein product.

Gene	Amino Acid			SIFT	Polyphen	PMut	Outcome
Position	Position	Red JF	AA				
1398	49	A	V	n/t [1]	benign	n/d [2]	n/d [2]
1422	57	N	K	n/t [1]	benign	n/d [2]	n/d [2]
1437	62	N	T	n/t [1]	probably damaging	n/d [2]	deleterious
1437	62	N	I	n/t [1]	benign	pathological	n/d [2]
1460	70	D	Y	n/t [1]	probably damaging	n/d [2]	deleterious
1460	70	D	H	n/t [1]	probably damaging	n/d [2]	deleterious
1464	71	Y	S	n/t [1]	probably damaging	pathological	deleterious
1464	71	Y	F	n/t [1]	possibly damaging	neutral	n/d [2]
1464	71	S	F	n/d [2]	benign	neutral	neutral
1466	72	G	E	n/t [1]	possibly damaging	pathological	deleterious
1509	86	R	K	tolerated	benign	neutral	neutral
1538	96	I	F	n/t [1]	probably damaging	n/d [2]	deleterious
1538	96	I	V	tolerated	benign	neutral	neutral

Red JF refers to the allele present in the reference genome sequence. AA stands for the alternative alleles. D70 is a catalytic site. N62 is in the catalytic cleft. [1] Not Tolerated. [2] Not determined. Site 49 is red, catalytic site 53 is black, catalytic cleft site 62 is yellow, and site 71 is blue.

Figure 6.5. A multiple sequence alignment of the active portion of the gene.



	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65		66	67	68	69	70	71	72	73	74	75	76	77
Domestic Chicken	K	V	F	G	R	C	E	L	A	A	A	M	K	R	H	G	L	D	N	Y	R	G	Y	S	L	G	N	W	V	C	V	A	K	F	E	S	N	F	N	T	Q	A	T	N	R	N	T	-	D	G	S	T	D	S	G	I	L	Q	I	N
Zebra finch	K	I	I	P	R	C	E	M	V	K	I	L	R	Q	N	G	F	Q	G	F	E	G	T	T	V	A	D	W	M	C	L	V	K	H	E	S	D	Y	N	T	K	A	Y	N	D	N	G	-	-	P	S	R	D	Y	G	I	F	Q	I	N
Turkey	K	V	Y	G	R	C	E	L	A	A	A	M	K	R	L	G	L	D	N	Y	R	G	Y	S	L	G	N	W	V	C	A	A	K	F	E	S	N	F	N	T	H	A	T	N	R	N	T	-	D	G	S	T	D	Y	G	I	L	Q	I	N
Copper pheasant	K	V	Y	G	R	C	E	L	A	A	A	M	K	R	L	G	L	D	N	F	R	G	Y	S	L	G	N	W	V	C	A	A	K	F	E	S	N	F	N	T	H	A	T	N	R	N	T	-	D	G	S	T	D	Y	G	I	L	Q	I	N
Kajil Pheasant	K	V	Y	G	R	C	E	L	A	A	A	M	K	R	L	G	L	D	N	Y	R	G	Y	S	L	G	N	W	V	C	A	A	K	Y	E	S	N	F	N	T	H	A	T	N	R	N	T	-	D	G	S	T	D	Y	G	I	L	Q	I	N
Golden Pheasant	K	V	Y	G	R	C	E	L	A	A	A	M	K	R	L	G	L	D	N	Y	R	G	Y	S	L	G	N	W	V	C	A	A	K	F	E	S	N	F	N	T	H	A	T	N	R	N	T	-	D	G	S	T	D	Y	G	I	L	Q	I	N
Bobwhite Quail	K	V	F	G	R	C	E	L	A	A	A	M	K	R	H	G	L	D	N	Y	R	G	Y	S	L	G	N	W	V	C	A	A	K	F	E	S	N	F	N	S	Q	A	T	N	R	N	T	-	D	G	S	T	D	Y	G	V	L	Q	I	N
Japanese Quail	K	V	Y	G	R	C	E	L	A	A	A	M	K	R	H	G	L	D	K	Y	Q	G	Y	S	L	G	N	W	V	C	A	A	K	F	E	S	N	F	N	T	Q	A	T	N	R	N	T	-	D	G	S	T	D	Y	G	I	L	Q	I	N
Human	K	V	F	E	R	C	E	L	A	R	T	L	K	R	L	G	M	D	G	Y	R	G	I	S	L	A	N	W	M	C	L	A	K	W	E	S	G	Y	N	T	R	A	T	N	Y	N	A	G	D	R	S	T	D	Y	G	I	F	Q	I	N
	•	:			•	•	•	:	.			:	:	:		•	:	:		:	.	•		:	:	.	:	•	:	•		.	•		•	•	.	:	•	:	:	•		•		•					•		•		•	:	:	•	•	•

	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137
Domestic Chicken	S	R	W	W	C	N	D	G	R	T	P	G	S	R	N	L	C	N	I	P	C	S	A	L	L	S	S	D	I	T	A	S	V	N	C	A	K	K	I	V	S	D	G	N	G	M	S	A	W	V	A	W	R	N	R	C	K	G	T	D
Zebra finch	S	K	Y	W	C	N	D	G	K	T	S	G	S	K	N	A	C	K	I	S	C	S	K	L	Q	D	D	N	L	G	D	D	I	Q	C	A	K	K	I	A	R	E	A	H	G	L	S	P	W	Y	G	W	K	N	H	C	R	G	K	D
Turkey	S	R	W	W	C	N	D	G	R	T	P	G	S	K	N	L	C	N	I	P	C	S	A	L	L	S	S	D	I	T	A	S	V	N	C	A	K	K	I	A	S	G	G	N	G	M	N	A	W	V	A	W	R	N	R	C	K	G	T	D
Copper pheasant	S	R	W	W	C	N	D	G	R	T	P	G	S	R	N	L	C	N	I	P	C	S	A	L	L	S	S	D	T	I	A	S	V	N	C	A	K	K	I	V	S	D	G	N	G	M	N	A	W	V	A	W	R	K	R	C	K	G	T	D
Kajil Pheasant	S	R	W	W	C	N	D	G	K	T	P	G	S	R	N	L	C	H	I	P	C	S	A	L	L	S	S	D	I	T	A	S	V	N	C	A	K	K	I	V	S	D	G	N	G	M	N	A	W	V	A	W	R	N	R	C	K	G	T	D
Golden Pheasant	S	R	W	W	C	N	D	G	R	T	P	G	S	R	N	L	C	H	I	P	C	S	A	L	L	S	S	D	I	T	A	S	V	N	C	A	K	K	I	V	S	D	G	N	G	M	N	A	W	V	A	W	R	N	R	C	K	G	T	D
Bobwhite Quail	S	R	W	W	C	N	D	G	K	T	P	G	S	R	N	L	C	N	I	P	C	S	A	L	L	S	S	D	I	T	A	T	V	N	C	A	K	K	I	V	S	D	G	N	G	M	N	A	W	V	A	W	R	N	R	C	K	G	T	D
Japanese Quail	S	R	W	W	C	N	D	G	R	T	P	G	S	R	N	L	C	N	I	P	C	S	A	L	L	S	S	D	I	T	A	S	V	N	C	A	K	K	I	V	S	D	V	H	G	M	N	A	W	V	A	W	R	N	R	C	K	G	T	D
Human	S	R	Y	W	C	N	D	G	K	T	P	G	A	V	N	A	C	H	L	S	C	S	A	L	L	Q	D	N	I	A	D	A	V	A	C	A	K	R	V	V	R	D	P	Q	G	I	R	A	W	V	A	W	R	N	R	C	Q	N	R	D
	•	:	:	•	•	•	•	•	:	•	.	•	:		•		•	:	:	.	•	•		•		.	.	:					:		•	•	•	:	:	.				:	•	:		.	•		.	•	:	:	:	•	:	.		•

	138	139	140	141	142	143	144	145	146	147
Domestic Chicken	V	Q	A	W	I	R	G	C	R	L
Zebra finch	L	S	S	F	V	R	G	C	-	-
Turkey	V	H	A	W	I	R	G	C	R	L
Copper pheasant	V	N	A	W	T	R	G	C	R	L
Kajil Pheasant	V	S	V	W	T	R	G	C	R	L
Golden Pheasant	V	N	A	W	T	R	G	C	R	L
Bobwhite Quail	V	Q	A	W	I	R	G	C	R	L
Japanese Quail	V	N	A	W	I	R	G	C	R	L
Human	V	R	Q	Y	V	Q	G	C	G	V
	:			:		:	•	•		

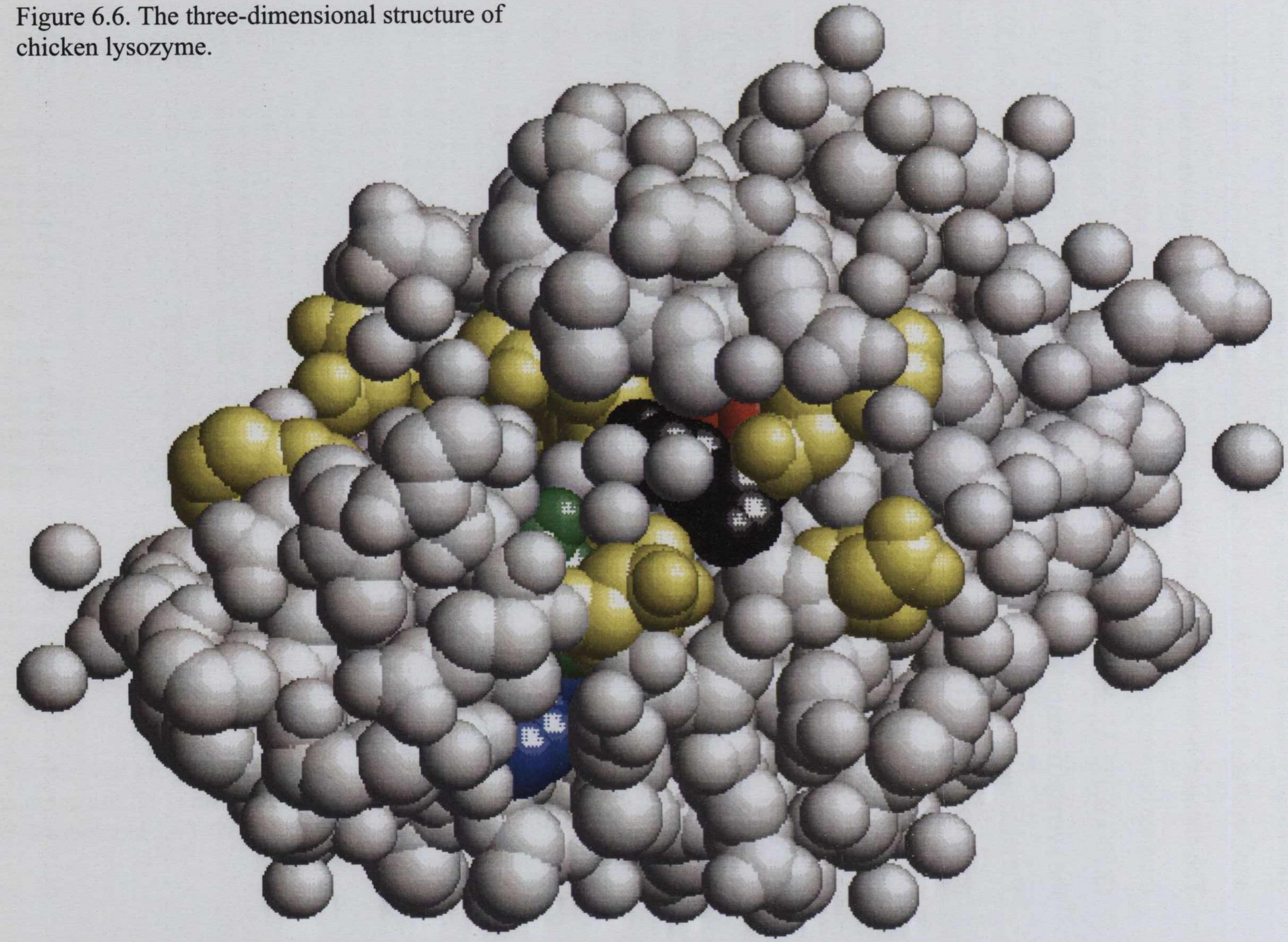
The cleaved pro-peptide sequences are not widely sequenced. The colours of the sites correspond to those in Figure 6: 49 is red, 53 is black, 70 is green and 71 is blue, and the catalytic cleft sites are in yellow (52, 55, 62, 75, 77, 80, 116, 119, 125, 126). Sites 19-53 and 103-149 are in the α-domain and 54-102 the β-domain (Blake *et al.* 1965, Schwalbe *et al.* 01).

6.3.4 Spatial relationship of variable and catalytic sites:

The positions of the catalytic sites (53 and 70), catalytic cleft and polymorphic sites (49 and 71) clustered closely in a three-dimensional model of chicken lysozyme displayed in RasMol 2.7.4.2 (Figure 6.6). Using the Euclidean distance between the α-carbon atoms of each amino acid, the length between sites 49 and 53 was 5.58 Å, substantially smaller than average, only 4.3% of site pairs were closer. Sites 70 and 71 were separated by 3.78 Å: only 0.9% of pairs and 10.9% of adjacent pairs were closer. The distance between these particular sites were small in comparison to the average distance between all sites (18.80 ± 9.47 Å), the mean distance between adjacent sites (3.80 ± 0.002 Å), the distance between the catalytic sites (53 and 70: 12.84 Å), and the average distance between the sites in the catalytic cleft (Figure 6.7: 25.04 Å). Variable sites 49 and 71 are also within 8 Å of several sites in the catalytic cleft: 52 is proximal to 49, and 71 to both 75 and 77.

The proximity of these amino acids to the catalytic sites was likely to be of significance because single sites changes and interactions can alter the activity and stability of this enzyme (Klein-Seetharaman et al. 2002, Zhou et al. 2007), implying that mutations at sites 49 and 71 spatially affected the catalytic sites. Certain sets of amino acid substitutions at lysozyme have been shown to be compensatory, even though they were located at the core of the molecule (Wilson *et al.* 1992): this could be possible for variants at sites 49 and 71.

Figure 6.6. The three-dimensional structure of chicken lysozyme.



The protein is displayed using RasMol. The positions of A49 (red) and Y71 (blue) are shown relative to the catalytic sites (E53, black; D70, green) and the catalytic cleft in yellow (F52, N55, N62, Q75, N77, W80, I116, D119, A125, W126). A49 and D53 are in helix 2. Y71 is located in sheet strand 2. The PDB file is available at http://www.rcsb.org/pdb/ under ID 3B6L; Michaux *et al.* 2008).

6.4 Discussion

A key innate immune hydrolase, lysozyme, was re-sequenced in Asian, African and broiler chicken populations and in six outgroup samples, including red, grey, Ceylon and green JF. Summary statistics, phylogenetic networks and interspecies tests for selection indicated a signal of elevated diversity in chicken balanced around one nonsynonymous site (A49V) and an additional nonsynonymous site (Y71S) that divided chicken and JF. Spatial modelling of the protein's structure suggested that these mutations could affect the catalytic function of the enzyme.

6.4.1 Polymorphic despite conservation:

Although nucleotide diversity at the lysozyme gene is lower (3.56 per kb) than the average between the red JF genome reference and a broiler (5.28 per kb; International Chicken Genome Sequencing Consortium 2004), there was considerable diversity maintained at noncoding regions. AMOVA results and statistical tests of Fu's F_S and *Hd* indicated that this was preserved mainly within populations. The positive Tajima's *D* and very negative Fay and Wu's *H* suggest that much of this haplotype diversity was present at the mid- and high-frequency levels in the samples, consistent with a hypothesis of balancing selection acting on the gene.

This balanced allele structure was modulated by recombination, which has generated alleles that were preserved at non-coding positions. In coding regions no such SNPs appear to be maintained, further supporting the conserved nature of the protein-coding portion of the gene. Purifying selection may quickly eliminate singletons created by recombination, ultimately leading to the scarcity of variation observed at the gene.

A previous study on diversity in commercial chickens found a dearth of rare alleles (Muir et al. 2008), an observation repeated here. The European broilers had much less haplotype diversity, which may be a result of breeding or of proximity to the origins of diversity for chickens: the further away from Asia, the harder it is for additional migration and introgressions to enhance diversity.

Studies in which A49V has been reported show it as a GCC (A) to GTT (V) change (Jung et al. 1980; also see NP_990612, Moult et al. 1976). In that paper, site 49 was segregating as GCT and GTT in chicken; GCC alone in the red, grey and Ceylon JF; and GCT in the bamboo partridge, grey francolin and green JF. A N124S mutation was also observed in the same analysis (Jung et al. 1980), which is also present in quail and pheasant, hinting that further coding sequence variability may exist at lysozyme – resequencing in more JF and bird samples may uncover additional polymorphisms of relevance to the functional diversity in commercial lines.

A49V was segregating in all eight sampled chicken populations, an indication that it was actively being maintained at high frequencies in each. Site 71 was different in chickens (S) compared to the sequenced red JF genome and all other birds (Y), and adjacent sites 70 and 72 seemed to be subject to positive selection between species. Unlike V49, S71 had risen to fixation in all observed chicken populations, bar one genotype that was F71. Interestingly, site 70 appears to have changed in green JF and grey francolin, perhaps a consequence of continued selection pressure at that site.

The noticeable absence of changes at nonsynonymous or synonymous sites suggested A49V and Y71S were likely to have function relevance. Calculations and visualisations of the distances between sites 49 and 71 to catalytic sites 53 and 70 indicated that alterations at 49 and 71 might cause changes to the enzymatic activity of lysozyme. The probability of one amino acid being substituted for another is dependent on several competing biochemical factors: size, pH and hydrophobicity (to a lesser also chemical composition and polarity; Conant 2009). These may need to be considered more carefully for lysozyme in the context that the chance of a mutation in an internal protein site is much lower than for an external one because internal changes may affect more adjacent sites. Lysozyme could therefore represent an anomaly among enzymes.

6.4.2 Identifying the cause of the balanced signal:

Though it is possible that the substitutions of interest here (A49V and Y71S) were a result of drift, founding effects or a build-up of deleterious mutations subsequent to domestication, like in dogs (Cruz et al. 2008), in light of the extensive conservation at the gene, this seems unlikely. It is more feasible that the diversity among the chicken populations distributed around substitution A49V may be a result of latent admixture following domestication and ongoing selective processes stimulated by novel pathogenic challenges.

The signature of high allelic diversity observed here is reminiscent of previous work on chicken genetic variation: mtDNA (Liu et al. 2006), MHC-B (Worley et al. 2008, O'Neill *et al.* 2009), Mx (Seyama et al. 2006, Berlin et al. 2008), IL1B and IL4RA, signifying that it may be the result of the complex population history of the chicken during domestication. Although the main source of chicken genetic variation is the red JF (International Chicken Genome Sequencing Consortium 2004, Fumihito et al. 1994), multiple domestications (Liu et al. 2006, Fumihito et al. 1996) and genetic introgressions of other JF into chicken populations (Silva et al. 2008, Nishibori et al. 2008, Eriksson et al. 2008) suggest diverse alleles may have been introduced during domestication. The impact of human trade, migration and selection for novel characteristics is likely to have created a widespread intermixing of red JF subspecies with chicken the result of which may be the high haplotype diversity observed in many studies (Oka et al. 2007, Granevitze *et al.* 2007, Kanginakudru et al. 2008, Muchadeyi et al. 2008, Bao et al. 2008, Berthouly et al. 2009). Though the elevated allele variation may be a relic of chicken domestication, this does not exclude the proposal of pathogen-driven selective pressure, which might explain the continued persistence of the divergent alleles in modern chicken populations.

6.5 Conclusion

A pattern of high allelic diversity observed in the resequencing of the lysozyme gene in global village chickens, broilers, jungle fowl, bamboo partridge and grey francolin was structured in a balanced manner around a replacement mutation at site 47 for the chicken samples. Broilers showed only slightly less relative variability than chickens from different countries in Asia and Africa. Variation differentiated between chicken and the other birds, including JF, by a replacement mutation at site 71, which is adjacent to catalytic site 70 in the enzyme. Red JF was the most likely candidate as the genetic resource for diversity at his gene.

Spatially aligning sites 47 and 71 showed they were sufficiently close to the two catalytic sites (53 and 70) to possibly alter their function. This suggestion of functional relevance was further highlighted by the paucity of coding sequence substitutions at the gene. Given that noncoding regions of the gene added to the balanced signal, it is likely that the admixture of populations from different domestication centres has enhanced variation in multi-modal fashion at many chicken genes, including lysozyme. Nonetheless, the maintenance of a pattern of diversity balanced around A49V at lysozyme indicates that it may still be subject to selection induced by pathogens.

Publication

This chapter formed the basis for a publication in Animal Genetics in 2009 entitled "Variation in chicken populations may affect the enzymatic activity of lysozyme". The authors are: Downing T, O'Farrelly C, Bhuiyan AK, Silva P, Naqvi AN, Sanfo R, Sow RS, Podisi B, Hanotte O and Bradley DG.

CHAPTER 7

The differential evolutionary dynamics of chicken cytokine and toll-like receptor gene classes

7.1 Introduction

The innate immune system provides an initial barrier against invasion and is initiated by pathogen recognition receptors (PRRs), including toll-like receptors (TLRs; Akira et al. 2000). The TLR family of transmembrane PRRs are key activators and regulators of immune response mechanisms that recognise pathogen-associated molecular patterns (PAMPs; Zhou et al. 2007). TLR extracellular domains identify conserved molecular moieties that are common to pathogens including lipopolysaccharide, flagellin, lipoproteins and microbial forms of nucleic acids. These detector molecules activate pathways through the TLR cytoplasmic domain. These signals are relayed through TLR pathway cascades activating the transcription factor NFκB to initiate expression of genes that code for molecules that amplify, mediate and regulate subsequent inflammatory and immune mechanisms, including cytokines (Leulier & Lemaitre 2008).

Cytokines are important immune mediators responsible for initiating, amplifying and regulating inflammation as well as controlling immune cell differentiation and proliferation in response to pathogenic challenge (O'Garra 1998). Chicken cytokines share properties with those in mammals (Staeheli et al. 2001), in which they perform equally extensive arrays of roles mediating innate and adaptive immune responses (Avery et al. 2004). Many chicken genes are orthologous to well-characterised mammalian cytokine class members, indicating that the chicken orthologs are likely to possess a wide range of functions (Kaiser et al. 2005). Their roles span the innate immune response initiated through IL1B and IL6 as well as the adaptive immune component. The latter has several cytokine groups defined by their roles in immunity (Kaiser et al. 2005): the cell-mediated T_H1 (IL12A, IL18 and IFNG); the humoral T_H2 (IL4 and IL13); and the anti-inflammatory (IL10 and TGFB). IL4 and IL13 mediate helper activities of T lymphocytes including differentiation of B cells. In contrast to mammals, IL5 expression is decreased during the T_H2 response (P. Kaiser, personal communication).

7.1.1 Cytokine and TLR genes associated with diseases:

Both cytokine and TLR classes selected for sequencing include genes implicated in response to parasitic, bacterial or viral diseases in avian species. To illustrate, TLR7 expression is increased but that of IL4 is suppressed during infection with avian influenza H9N2 (Xing et al. 2008), and PAMPs acting as agonists for chicken TLR2A, TLR4 and TLR5 can induce expression of IL8 (Kogut et al. 2005c). Expression of TLR4 increased in response to *SE* serovar Typhimurium and *Campylobacter jejuni*, but TLR5, TLR15 and IL8 (as well as IFNG) responded only to the former and not the latter (Shaughnessy et al. 2009). All nine TLRs resequenced here have been associated with chicken bacterial and viral diseases. Expression of TLRs increased in response to *SE* serovar Enteritidis, but that of TLR5 was downregulated (Abasht et al. 2008) – variation at this locus can affect resistance to *SE* serovars Enteritidis (Keestra et al. 2008) and Typhimurium (Iqbal et al. 2005). TLR7 showed elevated expression after infection with the former serovar, as did many pathway components activated by TLR signalling (Chiang et al. 2008). Nucleotide variation at TLR4 was implicated in resistance to diseases in commercial broilers (Keestra et al. 2009, Ye et al. 2006, Malek et al. 2004). This gene was associated with the immune response to challenges by gram negative *SE* serovar Enteritidis (as is TLR2A; Abasht et al. 2009), gram-positive Staphylococcus aureus (Farnell et al. 2003) and on exposure to PAMPs such as lipopolysaccharides (Ozoe et al. 2009, Keestra & van Putten 2008, He et al. 2006, Dil & Qureshi 2002). TLR3 was implicated in the immune response to influenza H5N1 infection (Karpala et al. 2008) and its expression and that of TLR2A was increased when exposed to the Massachusetts strain of infectious bronchitis virus (Wang et al. 2006). TLR15 and TLR2A expression was upregulated in response to *SE* serovar Typhimurium infection (Higgs et al. 2006). TLR15 was expressed more highly in chickens resistant to stimulation by *SE* serovar Enteritidis than those that were susceptible (Nerren et al. 2009). TLR1LA, TLR1LB, TLR2A and TLR2B were involved in activating the immune response to *Mycobacterium avium* (Higuchi et al. 2008).

Among the cytokines resequenced in this study, at least granulocyte-macrophage colony-stimulating factor (GMCSF), IL4, IL8, IL12 and IL13 have been linked with disease resistance or susceptibility. Expression of GMCSF, IL4 and IL13 was upregulated in response to Marek's disease virus infection (Heidari et al. 2008). IL8 expression was increased in response to transformation of cells with Rous sarcoma virus (Bedard et al. 1987, Sugano et al. 1987) as well as *Eimeria maxima* oocysts: the latter PAMPs also changed the expressions of the cytokine IFNG, IL1B, IL6, IL12, IL15, IL17A, IL10 and IL17D (Kim et al. 2008b). IL13 expression was increased following vaccination with the turkey herpes virus (Abdul-Careem et al. 2008). Higher expression levels of IL8 and IL12 were observed in chickens with better resistance to *Eimeria maxima* infection (Kim et al. 2008a). Expression of IL12 was elevated upon infection with four different *SE* serovars (Berndt et al. 2007). Several other cytokines not sequenced in this Chapter have also strong evidence for associations with diseases: pro-inflammatory cytokines IL1B, IL6 and IL18 respond to *Salmonella minnesota* LPS (Kogut et al. 2005b). *SE* serovar Enteriditis agonists stimulate these pro-inflammatory cytokines and also those involved in the T_H1 response: expression of IFNG in particular is important at all stages of the immune reactions (Kogut et al. 2005c). IFNG expression is also upregulated in environments with poor hygiene, as is that of IL2 (Ye et al. 2006). Additionally, IL1B and IL6 are activated by bacterial PAMPs via the TLR signaling pathway (Kogut et al. 2006).

7.1.2 Comparing cytokine and TLR gene classes:

The material presented in this chapter examines the variation present in two chicken immune gene classes from different functional categories to illuminate the adaptive pressures provided by domestication and disease. Genes whose products interact directly with the environment are more likely to have undergone adaptive change than those that mediate the immune response (Kim et al. 2007, Cui et al. 2009). Thus PRRs like TLRs are good candidates for detecting selection at the population level: previous studies have shown that the avian TLR genes have been subject to selection (Yilmaz et al. 2005, Cormican et al. 2009). Chicken cytokine genes may also be subject to selective processes, though the signatures may be more subtle (for example, IL1B in Downing et al. 2009a).

Recent advances in sequencing power have greatly enhance the potential for studies investigating genetic diversity (for example, Ng et al. 2009) and consequently, in this chapter Solexa high-throughput sequencing technology was used to examine variation at two categories of chicken immune genes with different functional roles: receptors and mediators. TLRs identify and alert the immune system to pathogen molecules, and cytokines act as mediators in regulating and communicating immune response signals. Nine genes from each class were resequenced in a panel of commercial and heritage chickens as well as red, grey, Ceylon and green JF (cytokine genes: IL3, IL4, IL5, IL8, IL9, IL12A, IL13, KK34 and GMCSF; and TLR genes: TLR15, TLR1LA, TLR1LB, TLR2B, TLR2A, TLR3, TLR4, TLR5 and TLR7).

Analyses indicated that a general pattern of high variability at these genes is likely to have been enhanced by genetic exchange between chicken and red JF, and in two possible instances between chicken and grey JF. Tests on variation and summary statistics between the gene classes indicated that the selection signatures at each gene class were distinctive: TLRs showed evidence of directional selection and cytokines evidence of diversifying selection. This difference was present in the allele frequency spectra at coding sites, suggesting functional relevance. The unique patterns of variation at each gene class may be constrained by their functional roles in the immune system: TLRs identify pathogens and thus are required to adapt quickly in response to pathogen evolution, whereas cytokines interact with many molecules in mediating the power of immune response signals, and consequently respond to selective stimuli differently.

7.1.3 Examining variability at MC1R:

Pigmentation, metabolism and the immune response have all been subjected to novel selective processes since domestication and key genes associated with these processes are likely to have undergone adaptation in chicken (Andersson 2003): for example, the gene determining chicken leg colour (yellow-skin) is derived from grey JF (Eriksson et al. 2008). An additional determinant of fowl plumage is Pmel17, a membrane glycoprotein involved in eumelanosome development (Kerje et al. 2004). Genes implicated in colouration determination are over-represented in a survey of cranially-expressed chicken-zebra finch orthologs displaying accelerated evolution (Axelsson et al. 2007). The melanocortin-1 receptor gene (MC1R) determines plumage colour in mammals and birds (Andersson 2003) by encoding a pleiotropic G-protein coupled transmembrane receptor expressed on melanocytes that increases intracellular cAMP levels when bound by α-MSH (α-melanocyte-stimulating hormone; Kerje et al. 2003). This change activates tyrosinase, which raises eumelanin production in melanosomes, darkening nearby feathers (Takeuchi et al. 1996a, 1996b). In this Chapter, diversity at MC1R was also investigated and this identified known plumage-associated mutations distinguishing chicken from JF.

7.2 Methods

7.2.1 Sample collection:

A total of 15 chicken samples were acquired: six Plymouth Rock heritage chickens (VIDO, Canada) and nine commercial broilers (Manor Farms, Co. Monaghan, Ireland) – five of the latter were Ross breed from Ireland and four were Hubbard Flex from France. The commercial broilers were a subset of those resequenced in Chapters 5 and 6. Nine jungle fowl (JF) were sampled: one green (CAS85707, Department of Ornithology & Mammalogy at the Californian Academy of Sciences); one Ceylon, one grey and one red (Wallslough Farm, Co. Kilkenny, Ireland); one Ceylon and one grey (Tommy Haran, Co. Meath, Ireland); and three red (Billy Wilson, Co. Antrim, Northern Ireland); a total of four red, two grey, two Ceylon and one green JF. DNA was isolated from the samples using a phenol-chloroform extraction following a proteinase K digestion.

7.2.2 Resequencing strategy:

The UCSC, GenBank and Ensembl genome resources were used to investigate the structures of the 19 genes (Figures 7.1-1 to 7.1-19). PCR primers were constructed using Primer3 software according to the parameters in Chapter 4 and were created by VHBio. The primer sequences' parameters were optimised for usage (Table 7.1). Each amplicon was amplified according to the PCR cycle setup (Table 7.2): 26 were successfully amplified – in most cases this included the entire coding region. All PCRs were performed with a Magnesium concentration of 20 mM. Not all amplicons successfully amplified for every sample.

Although a SNP-based study proposed that commercial chickens have a 50% deficit of rare alleles (Muir et al. 2008), this was not observed to the same extent in previous resequencing-based analyses that used commercial chickens included in this thesis. In addition, SNP-based approaches may be compromised by an ascertainment bias that misses many low-frequency variants (Kreitman & Di Rienzo 2004, Soldevila et al. 2005), which is reduced here by aiming to resequence entire genes at high coverage rates.

Table 7.1. Sets of primer pair sequences used in PCR and their optimal parameters.

Gene	Amplicon	Size (bp)	Orientation	T_M	Primer Sequences
GMCSF	1	2487	Forward	63	ATATAAGAGGAACACAGGGCAGAG
			Reverse		TGAGATTACAGCAGTGAAAGCAG
IL12A	1	1894	Forward	61	TCCTACTCCTCCACCGACATAA
			Reverse		CTGCGTTTGCTTCTTACATCTCT
IL13	1	1892	Forward	61	CAGCATTTTGACTGTAGTGAGCA
			Reverse		GTTGGCAAGCACTCTGGTTAT
IL3	1	2608	Forward	58	TGAGCTTCTTTGTGGTGAGGTAT
			Reverse		GTTCAGATGTGTCAACTCCCTCTA
	2	1533	Forward	59	ATTGACTCCAAGCCAAGTAAGTG
			Reverse		ATGGTTCCCCTCACTAAACAAAGT
IL4	1	1943	Forward	61	CTGCCTCCTACCACTGTTATCTG
			Reverse		GGTCTGCTAGGAACTTCTCCATT
IL5	1	1090	Forward	56	AGCAAACACTTGGATGTGACC
			Reverse		TTGGCTCTCAATAAAAGCTAGA
IL8	1	2524	Forward	61	AAACAAGCCAAACACTCCTAACC
			Reverse		CACAGCACTGACCATTATGAAAG
IL9	1	2609	Forward	61	GGACAATCCTGCTTTGAACTCT
			Reverse		CACGTGTCAGACTCTGGTAGAAG
KK34	1	1100	Forward	58	AGTTGACAGCTGAGAATGAAGACTCAC
			Reverse		ATTGGACACGCTGCCTTCAA
TLR15	1	3072	Forward	59	ATCCTTCTGACACCTCTTCTAGT
			Reverse		TGCAGTAATCTCCAAAAGATAGT
TLR1LA	1	2963	Forward	60	AGGTCACGTAGTCCAACTCTCTG
			Reverse		CAGCAATTTAGGAACGCTTCAC
TLR1LB	1	1447	Forward	60	GATGGGATCTGTGGAAGAGTAAAG
			Reverse		CATCTTGGGAAGGTCTAAGTATGG
	2	2513	Forward	60	GAATGTGCATTGTAGACCCAGTAG
			Reverse		GAATAGTCGAAGCGAGTACTTACG
TLR2B	1	3104	Forward	60	CTAATTCTCATCTGTTCCCAGCAC
			Reverse		ATACTGAAACGAGCTCCTAACCTG
	2	2761	Forward	61	TTCAGAAAGACAGAACAGCAAGG
			Reverse		TCCAGTAGAGGATGGCTACAGTC
TLR2A	1	5932	Forward	60	TGGCCTACACAGACATATTCTAGC
			Reverse		CAGTTGGAGTCGTTCTCACTGTAG
TLR3	1	3157	Forward	59	CAGTCTGCCTGATACTCTCACTTG
			Reverse		AGTGGTAGTGTCCATTCTCCTTTC
	2	2888	Forward	56	CCAGTCTGCCTGATACTCTCACT
			Reverse		GCAACCTTCAGTGACTTATTCCA
	3	2286	Forward	58	GTAACGGAGTCTCTTCACTCTGC
			Reverse		CCACACCATACTTCATCAGCATA
TLR4	1	2530	Forward	59	TTAGTGCGGTAGTGTTAGTGAAGG
			Reverse		GTTGCCACTCCTTATCTTGATAGC
	2	2066	Forward	59	ATTCCCCAACTCTACAGCTACATC
			Reverse		TGCACTCAGTATCTGGACTGAAAG
TLR5	1	3086	Forward	59	TTAAGCCAATGTACCAGAGTAGT
			Reverse		TTCCAAGTTTAGTAGGATTTTCA
TLR7	1	3396	Forward	60	GCTGCTGTTGTCTTGAGTGAGT
			Reverse		CAGAAATGAACGTGTAGGAAGGA
MC1R	1	2377	Forward	60	TTTGTAGGTGCTGCAGTTGTG
			Reverse		TGAATTGCAGATGATGAGGATG

Certain regions amplified are overlapping. T_M is the annealing temperature (°C).

Table 7.2. PCR cycle program used for each primer pair.

Step	T_M	Duration (min)
1	95	15
2	95	0.5
3	T_M	0.75
4	72	1
5	72	15

T_M is the annealing temperature listed in Table 7.1 (°C). Steps 2 to 4 were repeated 33 times in sequence.

7.2.3 Sequence assembly:

Equimolar PCR libraries were assembled for each individual chicken and JF sample. Each individual PCR library was coligated, tagged and then pooled into two library sets containing 12 PCR sample libraries each by GATC, UK (www.gatc-biotech.com) – one set of 12 PCR libraries for each of two lanes on a flow cell. Pooled PCR library preparation and resequencing on an Illumina Solexa Genome Analyser II was carried out by GATC. Using efficient local alignment of nucleotide data software (Eland; Cox, unpublished), mapping of sequences to reference genes from GenBank (Table 7.3) allowed a pair of 36-base reads to be clustered where there were not more than 2 mismatches over their lengths.

Table 7.3. Average coverage values for each gene from Solexa resequencing.

Gene	Coverage	GenBank accession number
GMCSF	444.4	NM_001007078
IL12A	232.0	NM_213588
IL13	241.0	NM_001007085
IL3	229.4	NM_001007083
IL4	693.8	NM_001007079
IL5	230.4	NM_001007084
IL8	384.2	NM_205498
IL9	320.3	NM_001037825
KK34	211.1	NM_213585
TLR15	205.7	NM_001037835
TLR1LA	85.5	NM_001007488
TLR1LB	425.8	NM_001098854
TLR2B	350.7	XM_001232192
TLR2A	56.7	NM_204278
TLR3	432.0	NM_001011691
TLR4	407.7	NM_001030693
TLR5	78.3	NM_001024586
TLR7	198.0	NM_001011688
MC1R	37.1	NM_001031462

A total of 22,607,104 reads were generated – equivalent to 814 Mb of DNA. This gave a mean of 1,189,848 ± 482,630 reads per gene, of which 77% on average aligned correctly to the reference gene sequences. Reads that did not align to these genes were either of poor sequence quality and so exceeded the number of mismatches per read, or aligned to control DNA regions inserted to test the robustness and fidelity of alignments. The mean coverage for all sequences was 290 ± 154: 332 ± 149 for cytokines and 249 ± 148 for TLRs (Table 7.3). A total of 55,848 bp of valid gene sequence was amplified – 19.7 kb for the cytokines and 36.2 kb for the TLRs. A total of 927 bp in two blocks spanning 2,377 bp was amplified at MC1R.

7.2.4 SNP ascertainment:

SNPs were called in the verified aligned sequences according to a set of criteria as follows:

(1) if the reference sequence was known and different to the resequenced data;

(2) if the base coverage > 20;

(3) if the fraction of undetermined nucleotides ("N") < 0.25;

(4) if the polymorphic nucleotide frequencies observed > 0.35 of the total, including heterozygotes;

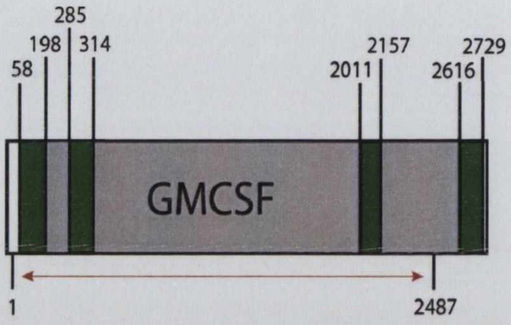
(5) if the base quality > 30, so the probability that the base is called correctly ≥ 0.999.

The base quality ranged from 0 to 100: most nucleotides had base quality ≥ 99. All likely SNPs were verified visually: 77 of them (8.3%) clustered as serial SNPs suggestive of localised poor sequence quality and were omitted from analysis.

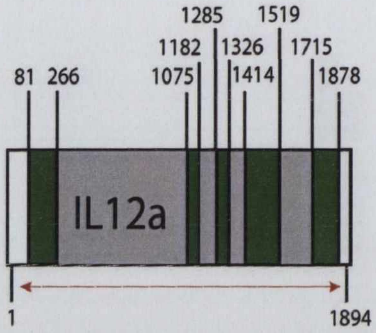
The sequences for all these genes have accession numbers in GenBank: FJ907553-600 for GMCSF, FJ907601-48 for IL3, FJ907649-96 for IL12A, FJ907697-742 for IL13, FJ907743-90 for IL4, FJ907791-838 for IL5, FJ907839-82 for IL8, FJ907883-924 for IL9, GQ337430-GQ337473 for KK34, FJ915219-58 for TLR15, FJ915259-98 for TLR1LA, FJ915299-342 for TLR1LB, FJ915343-86 for TLR2B, FJ915387-432 for TLR2A, FJ915433-80 for TLR3, FJ915481-528 for TLR4, FJ915529-54 for TLR5, FJ915555-600 for TLR7, and FJ915199 to FJ915218 for MC1R.

Figure 7.1. Structures of genes resequenced.

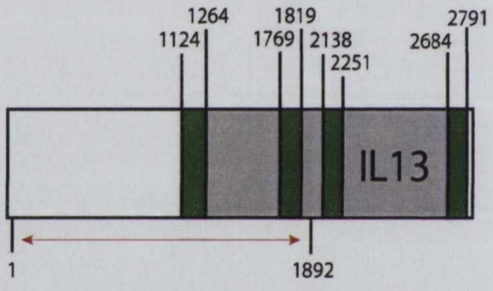
Gene names are listed in diagrams: 1 is GMCSF, 2 is IL12A, 3 is IL13, 4 is IL3, 5 is IL4, 6 is IL5, 7 is IL8, 8 is IL9, 9 is KK34, 10 is TLR15, 11 is TLR1LA, 12 is TLR1LB, 13 is TLR2B, 14 is TLR2A, 15 is TLR3, 16 is TLR4, 17 is TLR5, 18 is TLR7 and 19 is MC1R. Exons are in green, introns in grey, UTRs in blue and poor quality resequenced regions in black. Intergenic regions are shown as white and amplified regions are indicated by the red arrows. The numbers shown represent the base positions in relation to the GenBank gene sequences. The black region at MC1R has a very high GC content and consequently did not amplify adequately.



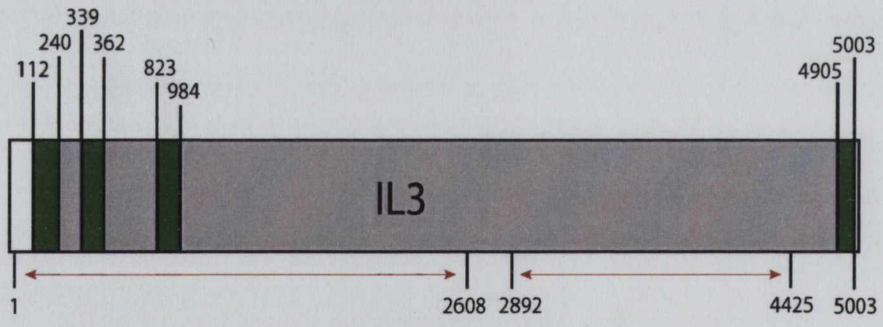
(1)



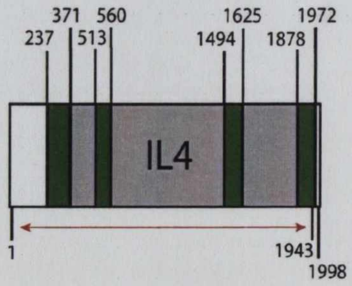
(2)



(3)

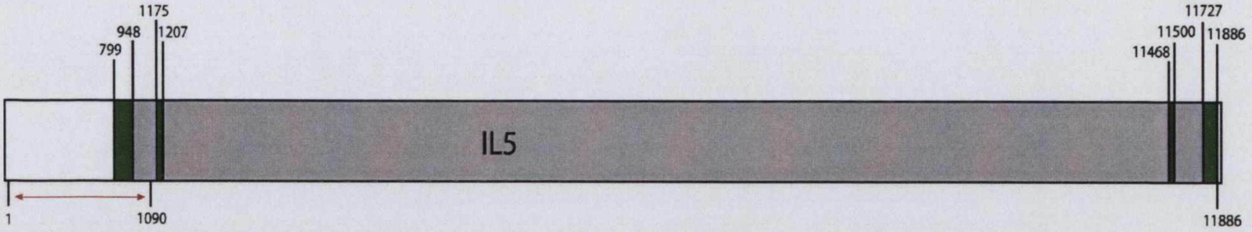


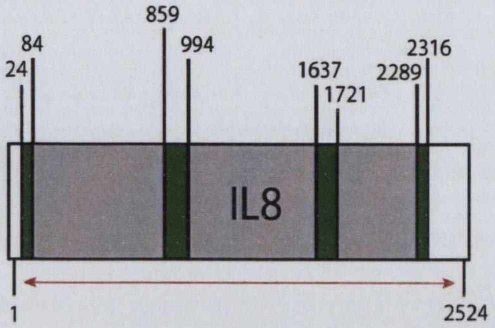
(4)



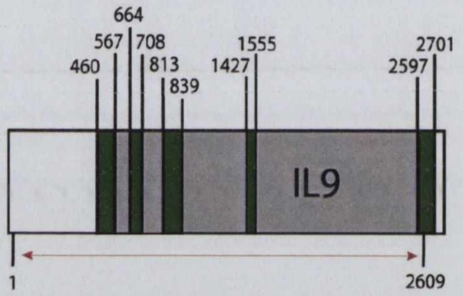
(5)

(6)

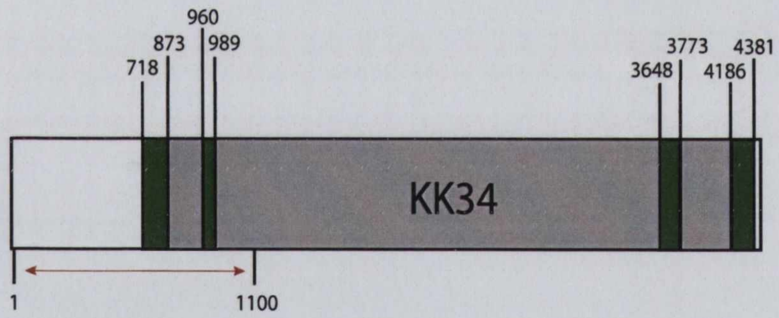

1175
948
1207
799
IL5
11727
11500
11886
11468
1
1090
11886


859
84
994
24
1637
2316
1721
2289
IL8
1
2524

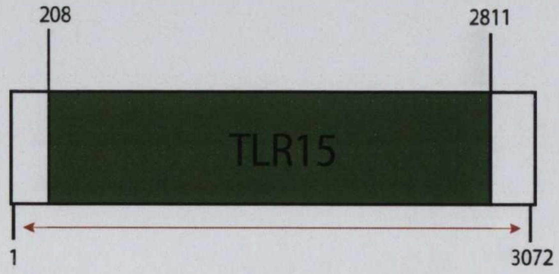
(7)


664
567
708
460
813
839
1555
1427
2701
2597
IL9
1
2609

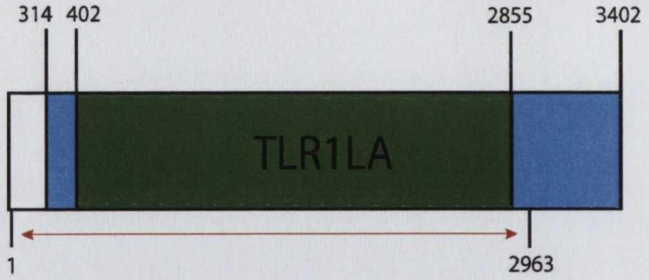
(8)


960
873
989
718
3773
4381
3648
4186
KK34
1
1100

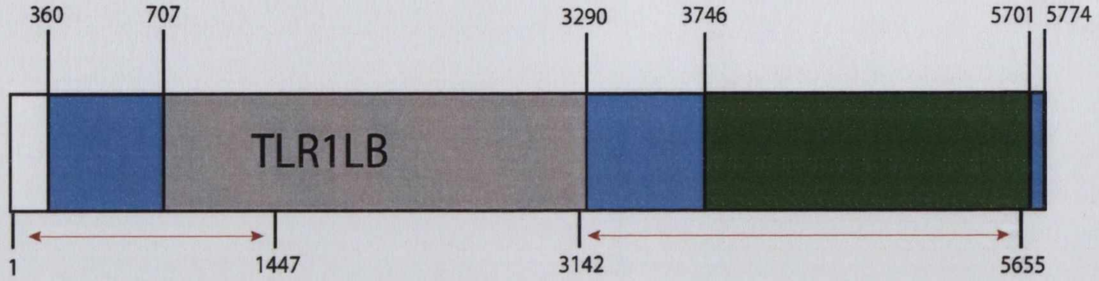
(9)


208
2811
TLR15
1
3072

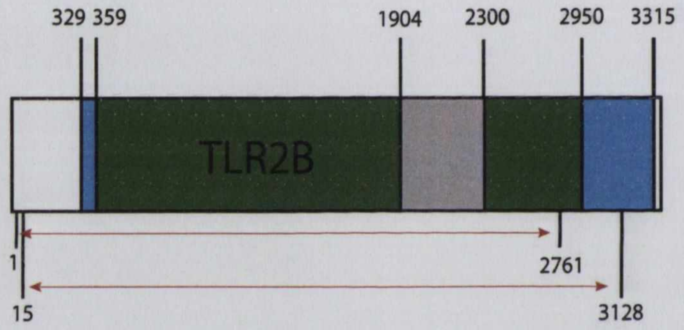
(10)


314
402
2855
3402
TLR1LA
1
2963

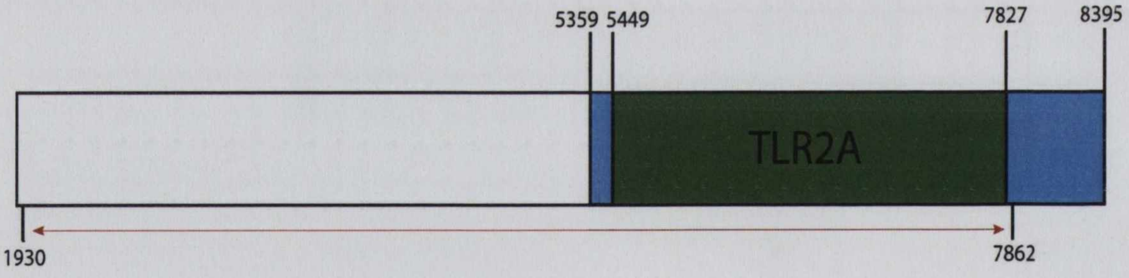
(11)


360
707
3290
3746
5701
5774
TLR1LB
1
1447
3142
5655

(12)

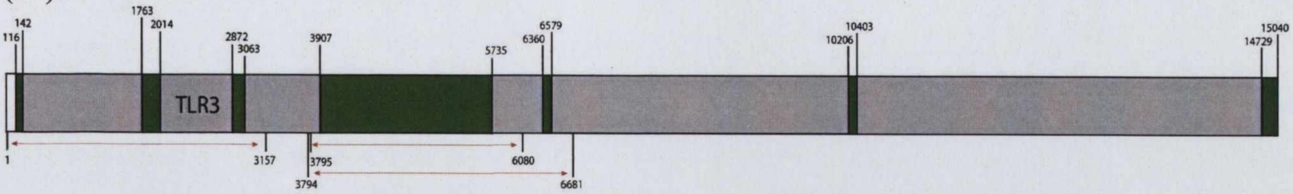

329
359
1904
2300
2950
3315
TLR2B
1
2761
15
3128

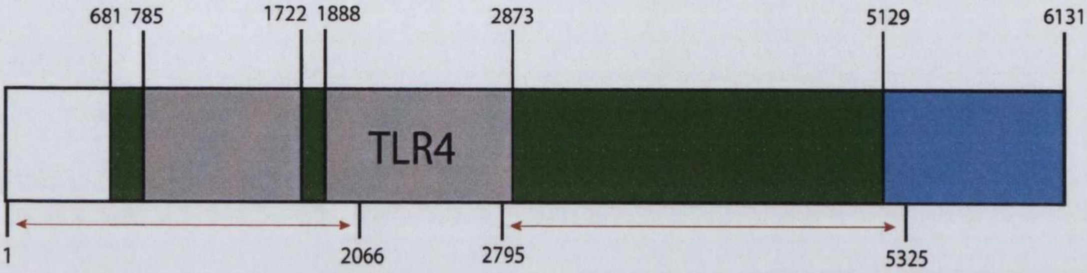
(13)


5359
5449
7827
8395
TLR2A
1930
7862

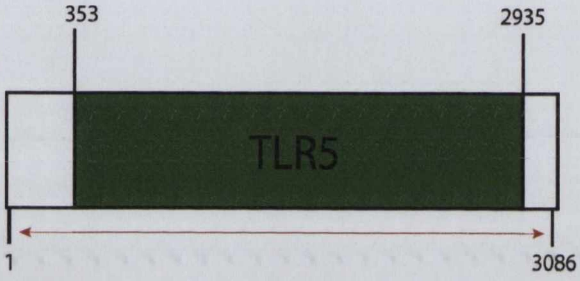
(14)

(15)

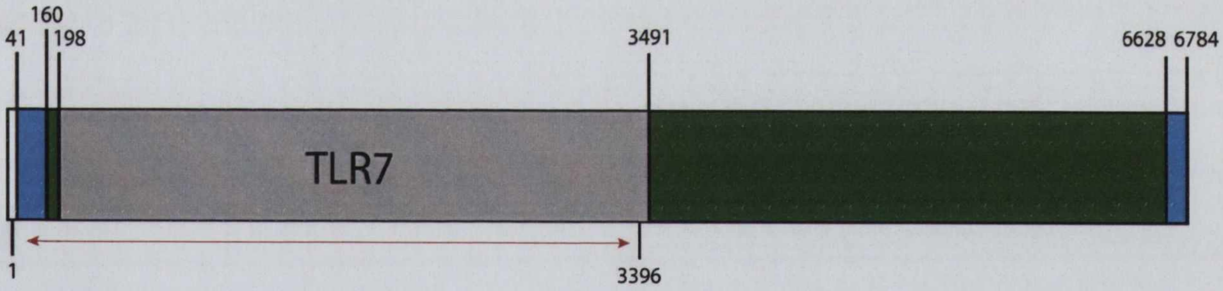

1763
142
2014
116
2872
3063
3907
5735
6579
6360
10403
10206
15040
14729
TLR3
1
3157
3795
6080
3794
6681


681
785
1722
1888
2873
5129
6131
TLR4
1
2066
2795
5325

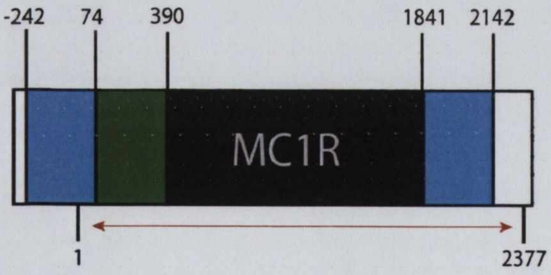
(16)


353
2935
TLR5
1
3086

(17)


160
41
198
3491
6628
6784
TLR7
1
3396

(18)


-242
74
390
1841
2142
MC1R
1
2377

(19)

7.2.5 Data analysis:

To determine the genetic relationship between chicken and each JF species, F_{ST} (Wright 1951) values for populations and species were tested for 1,000 permutations using Arlequin (Schneider et al. 2000). Additionally, median-joining haplotype networks were constructed for each gene using Network version 4.2.0.1 (Bandelt et al. 1999). The connection cost criteria for joining nodes was used for all genes except GMCSF where it did not converge on a single network, so a greedy algorithm (greedy FHP), which joins the nearest nodes at each iteration, was used. AMOVA analysis (Excoffier et al. 1992) was carried out on MC1R using Arlequin version 2.001 (Schneider et al. 2000) to quantify the extent of population and species differentiation.

DnaSP 5.0 (Librado & Rozas 2009) was used to analyse the numbers of SNPs and haplotypes, haplotype diversity (*Hd*, Depaulis & Veuille 1998) and Watterson's estimator of genetic diversity ($\theta_W = 4N_e\mu$; Watterson 1975). Nucleotide diversity was measured using π (Tajima 1983a). Nucleotide divergence between chicken and JF species was assessed using *K*, the average number of nucleotide differences per site. Values for π_A (the population average number of nonsynonymous SNPs per nonsynonymous site), π_S (the population average number of synonymous SNPs per synonymous site) and equivalent metrics for interspecies divergence (K_A and K_S) were also calculated. The four gamete test to get the minimum number of recombination events (R_M; Hudson & Kaplan 1985), *R* (the degree of recombination; Hudson 1987), GC content and gene conversion (Betran et al. 1997) were also analysed. π per kb, θ_W per kb and *K* per kb were adjusted for the poor quality region in MC1R (length 1,450 bp).

A set of summary statistics were used to identify departures from neutrality using coalescent simulations in DnaSP: Fu and Li's *D* and *F* (Fu & Li 1993), Tajima's *D* (Tajima 1983b), Fu's F_S (Fu 1996), Fay and Wu's *H* (Fay & Wu 2000). These tests were performed as detailed in Chapter 4. These simulations generated empirical distributions with which the observed values were compared to determine the extent of their deviation from neutrality. Non-neutral evolution was inferred if the observed values lay at the extremes of the distribution.

Deleterious alleles are expected to remain at low frequencies, whereas only functionally advantageous or neutral alleles are expected to rise to intermediate or high frequencies, particularly in large populations (Axellson & Ellegren 2009). The fraction of deleterious alleles can be estimated by examining derived allele frequencies (DAF). The fraction of alleles with DAF < 0.2 minus that for alleles with DAF > 0.2 (Liti et al. 2009) can approximate the deleterious proportion. This was carried out for DAF at amino acid-altering (the number of nonsynonymous changes per nonsynonymous site: DAF_N) and silent coding (the number of synonymous changes per synonymous site: DAF_S) sites. Variation at synonymous nucleotides is expected to be neutral, whereas that at nonsynonymous sites would be subject to selective processes (Yang 2002). The fraction of nonsynonymous substitutions expected to be neutral (*f*) was determined for each gene as well (Eyre-Walker and Smith 2002).

In order to determine the neutrality of population allele frequency spectra at functional sites in the gene classes, the DAF of SNPs at nonsynonymous (DAF_N) and synonymous (DAF_S) sites were determined. This corresponds to the DAF for each coding SNP site. These values were also determined for the entire sampled population (denoted π_A and π_S) where they were quantified in terms of their specific DAF relative (π_A and π_S) to the average DAF for all genes (DAF_N and DAF_S). Ancestral and derived alleles were determined according to those present in grey, Ceylon and green JF; the multiple origins of the domestic chicken complicate signatures of ancestry from red JF.

The protein domain locations of nonsynonymous mutations were ascertained from Uniprot. If there were no annotated chicken gene protein domains, these were determined by aligning the chicken genes with their human orthologs using T-Coffee (Notredame et al. 2000).

The sizes of the TLR protein extracellular, cytoplasmic and transmembrane domains were estimated using TMpred (http://www.ch.embnet.org/software/TMPRED_form.html; Hofmann & Stoffel 1993). This software compares the protein sequence to entries in SwissProt database 22 from which it can estimate the most probable transmembrane region and its cell membrane orientation, which serves to specify the domains. The sole major limitation of TMpred is its inability to estimate confidently signal peptide domains, which tend to be small in comparison to the TLR proteins' full lengths of about 700-1000 sites. This program was used because it does not rely purely on mammal-chicken protein alignments, which can be highly divergent in certain regions, and is effective for novel avian genes, such as TLR15, which have no mammalian ortholog.

The lengths of signal and active peptide regions of the cytokines were determined with SignalP 3.0, which ascertains the most likely cleavage points using the hidden Markov model methods (http://www.cbs.dtu.dk/services/SignalP/; Bendtsen et al. 2004). By comparing amino acid composition of N-terminal regions of the input protein to those of a eukaryotic database, the chances of each site belonging to the signal peptide and the cleavage region were calculated so that the amino acids with the maximum cleavage site probability could be identified.

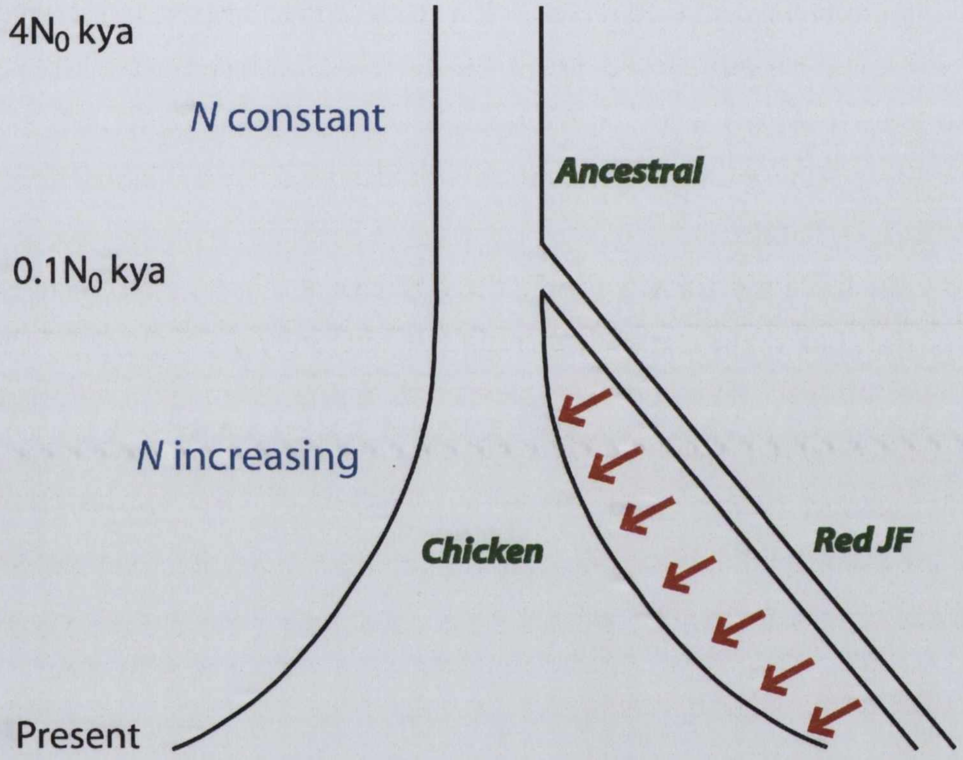
Using the codeml implementation of PAML 3.15 package (Yang 1997), as $\omega = d_N/d_S$ was calculated for the coding sequences of MC1R in chicken, red, grey and Ceylon JF where the d_N was the relative rate of nonsynonymous mutations and d_S the relative rate of synonymous mutations (Yang 2002). Branch-specific models determined one ω for the chicken lineage and another for the thee (grey, Ceylon and green) JF lineages according to a neutral model where $\omega = 1$ and a variable model where ω can vary (Yang 2002). A LRT was performed between the log likelihoods of the variable and neutral models according to a χ^2 distribution with one degree of freedom.

7.2.6 Simulating demographic history:

Coalescent simulations were used to test if chicken demographic history could explain the variation at the cytokine and TLR gene classes. Using the program MS (Hudson 2002), samples were generated under a neutral model for 100,000 repetitions given the observed resequenced data observed for each of the 18 genes (numbers of SNPs, θ_W, number of samples, recombination and gene length). Models simulated chicken population size growth and genetic introgression of JF into the chicken population so that comparisons between observed and simulated data would indicate if these two parameters affect diversity at these loci.

In order to simulate the original domestication of chicken from red JF and subsequent genetic introgression of red JF, the simulations started $4N_0$ kya with an ancestral mixed population of size N_0 (Figure 7.2). At $0.1N_0$ kya, the population divided in two groups, chicken and red JF – the population size of the latter was constant. The growth rate (α) of the present chicken population size (N_t) was determined as $N_t = N_0e^{-\alpha t}$: given that chickens have a large effective population size, $1 \leq \Delta N \geq 100$ for $\Delta N = N_t/N_0$ for t, the present time in generations (International Chicken Genome Sequencing Consortium 2004).

Figure 7.2. Demographic model for coalescent simulations.



After domestication $0.1N_0$ kya, a historically panmictic ancestral population with a constant size since $4N_0$ kya split into a chicken population, whose population size increased, and a red JF population, whose population size was static. Red JF genotypes migrated into the chicken population, represented here by the red arrows. The rates of chicken population size expansion and introgression of red JF genotypes were varied in coalescent simulations.

For a migration rate m, introgression was simulated as the replacement of red JF genotypes in the chicken population at $0 < 4N_0m < 100$ per kyr (one generation was taken as one year). Limiting the extant of introgression (up to a maximum of 0.25 per kyr) was based on the likelihood that higher values were unlikely to be realistic, given that these were likely to be sporadic events. Similarly, the present chicken population size (N_t) was limited to $100N_0$. Although F_{ST} can be used to estimate the migration rate between populations as $F_{ST} = (4N_e m + 1)^{-1}$ (Holsinger & Weir 2009), this approach did not converge on a single migration rate. Simulations were conducted for each of the 18 genes for three Tajima's D values: one neutral with $D = 0$, a second as the average cytokine D, and the third the mean TLR D. MS simulates the number of SNPs and π. Likelihoods of models with varying rates of population expansion and introgression were calculated in comparison to the observed data. LRTs compared the model with the maximum likelihood with alternative models according to a χ^2 distribution and thus determined ranges of growth and introgression values which were significantly more likely for each of the neutral, cytokine and TLR D values.

For the range of values such that $0 < 4N_0m < 100$ given $4N_0m$ = g/kyr and $1 < N_t/N_0 < 100$ where $N_t/N_0 = \Delta N$ were determined for a set of simulated values defined as x_1, x_2, … , x_{18}, representing each of the 18 genes. The likelihood ($log_e(L)$) was determined where δ is the gene sample standard deviation, x_i is a the i^{th} of 18 genes and μ is the gene sample mean:

$$\log_e(L) \cong -\frac{1}{2\delta^2}\left(\log_e\left(\delta\sqrt{2\pi}\right)^{-18}\right)\sum_{1}^{18}\left(x_i - \mu\right)^2$$

Constant introgression at a rate of $4N_0m$ will give the fraction of the current population that were from the domesticated chicken founding population according to $(4N_0m)^{10}$: a rate of introgression of 0.03 is equivalent to the present chicken population being 0.737 chicken and 0.263 red JF; 0.06 is equivalent to 0.539 chicken and 0.461 red JF; 0.09 to 0.389 chicken and 0.611 red JF; 0.12 to 0.279 chicken and 0.721 red JF; 0.15 to 0.197 chicken and 0.803 red JF; 0.18 to 0.137 chicken and 0.863 red JF; 0.21 to 0.095 chicken and 0.905 red JF; and 0.24 to 0.064 chicken and 0.936 red JF.

7.3 Results

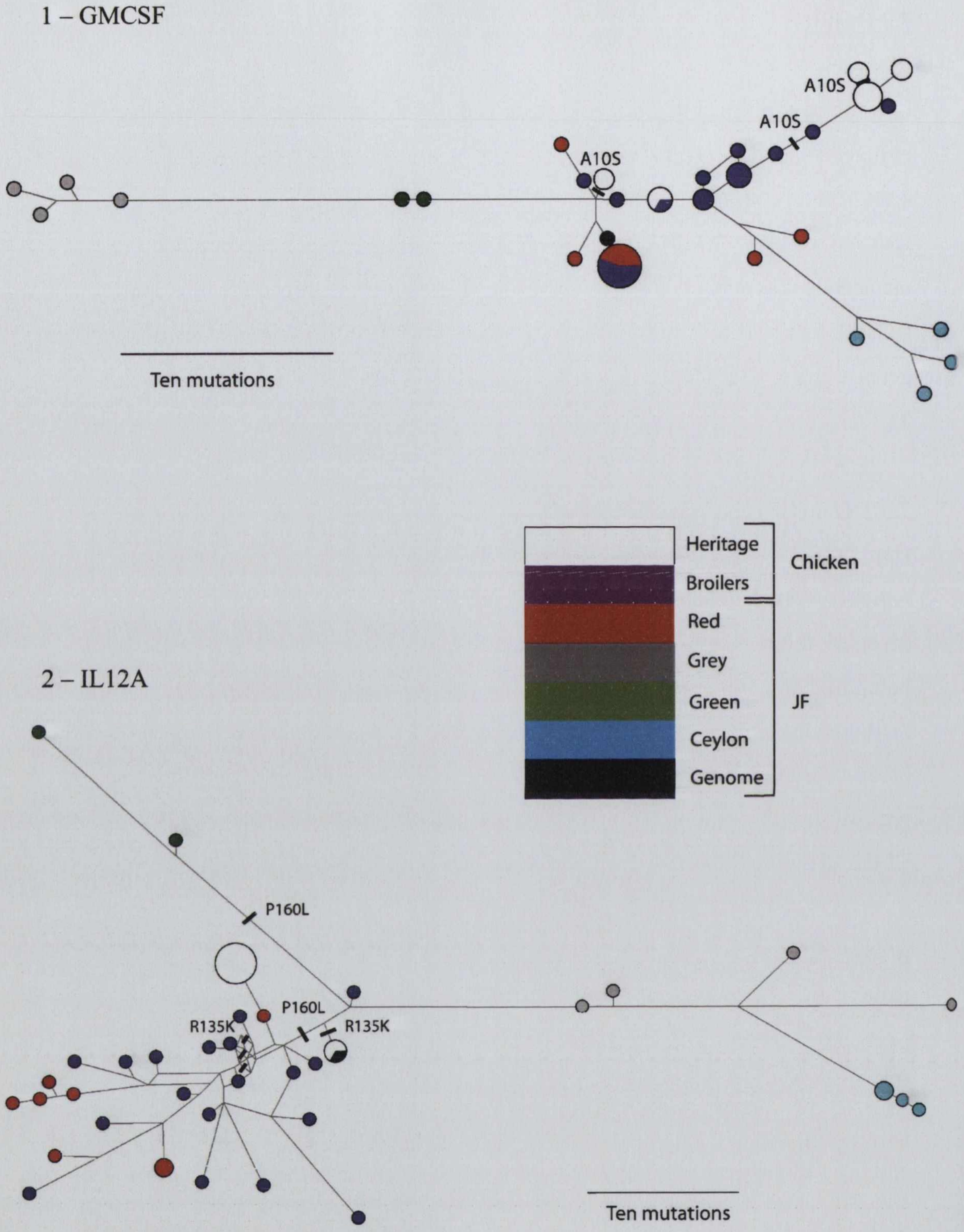
In order to explore both chicken genetic history and immune gene category evolution, two gene classes related to immune defence and the MC1R gene were resequenced in a set of nine commercial broilers, six heritage chickens and nine closely related JF: four red, two grey, two Ceylon and one green, with an average coverage of 290x using Solexa sequencing technology. The classes were a set of nine immune mediator (IL3, IL4, IL5, IL8, IL9, IL12A, IL13, KK34 and GMCSF) and nine immune receptor genes (TLR15, TLR1LA, TLR1LB, TLR2B, TLR2A, TLR3, TLR4, TLR5 and TLR7).

7.3.1 Genetic relationship between chicken and jungle fowl:

Haplotype phylogenetic networks showed no evidence of an ancient division of variation between chicken and red JF (Figures 7.3-1 to 7.3-18). At most genes diversity followed a trend of being lower in heritage birds than in broilers, and of being higher again in red JF. The chicken samples were phylogenetically distinguishable from grey, Ceylon and green JF, as were the JF species from each other – for example, TLR4 (Figure 7.3-16). However, there were two genes where grey JF shared variation with chicken and red JF: at TLR1LA (Figure 7.3-11) and TLR2A (Figure 7.3-14) – results of resequencing MC1R are in section 7.3.11.

Figure 7.3 (over leaf). Median-joining haplotype networks of chicken and JF for each gene.

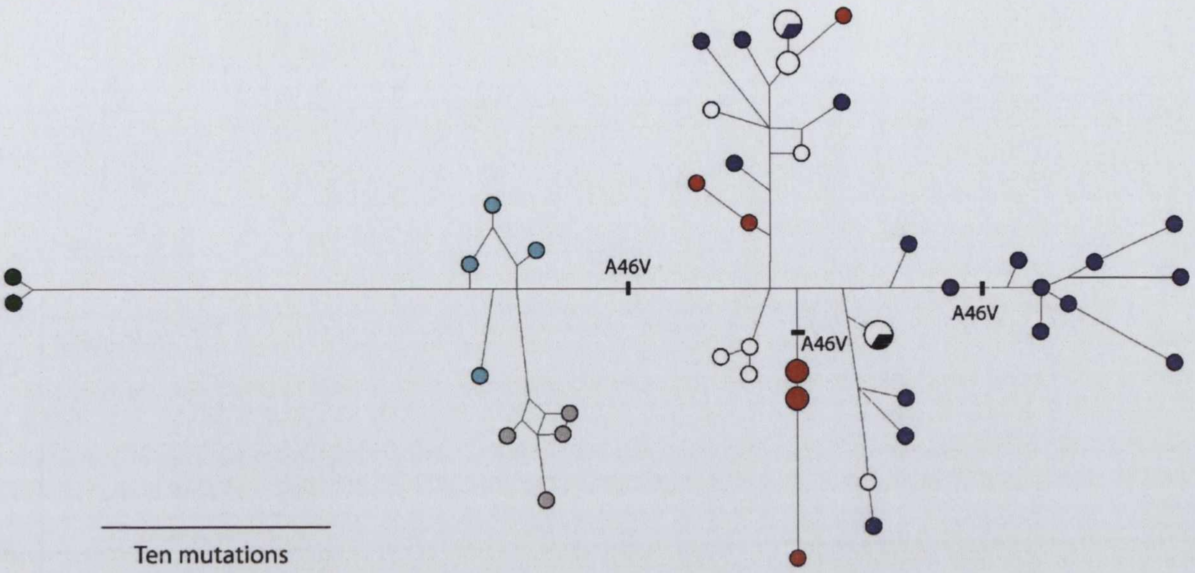
Labels: 1 is GMCSF, 2 is IL12A, 3 is IL13, 4 is IL3, 5 is IL4, 6 is IL5, 7 is IL8, 8 is IL9, 9 is KK34, 10 is TLR15, 11 is TLR1LA, 12 is TLR1LB, 13 is TLR2B, 14 is TLR2A, 15 is TLR3, 16 is TLR4, 17 is TLR5 and 18 is TLR7. Broilers are purple; heritage chickens are white; red JF are red; Ceylon JF are blue; green JF are green; and grey JF are grey. Branch lengths are proportional to the mutational distance listed and circle size is proportional to number of birds at each node. Disease-associated SNPs at TLR4 (L73, K83E, D301E, K343R, H383Y, R611Q) and nonsynonymous SNPs where the frequency in the chicken populations was intermediate or higher (≥ 0.1) are shown by black bars: K38Q, N44S for TLR3; A10S at GMCSF; R135K and P160L at IL12A; A46V at IL13; M23T and V49I at IL4; L7F and V12M at IL5; S19C, S32R and M34A at KK34; A309E at TLR15; V788A and C815R at TLR1LA; P38S, G45D, F99L, R106Q, D119N, V123I and I637V at TLR1LB; V196L at TLR2B; S516R at TLR2A; T280S, R345S, G362E, K459R, A540V, D545H, A592S and A649V at TLR3; R212K at TLR5 (see Table 7.10).



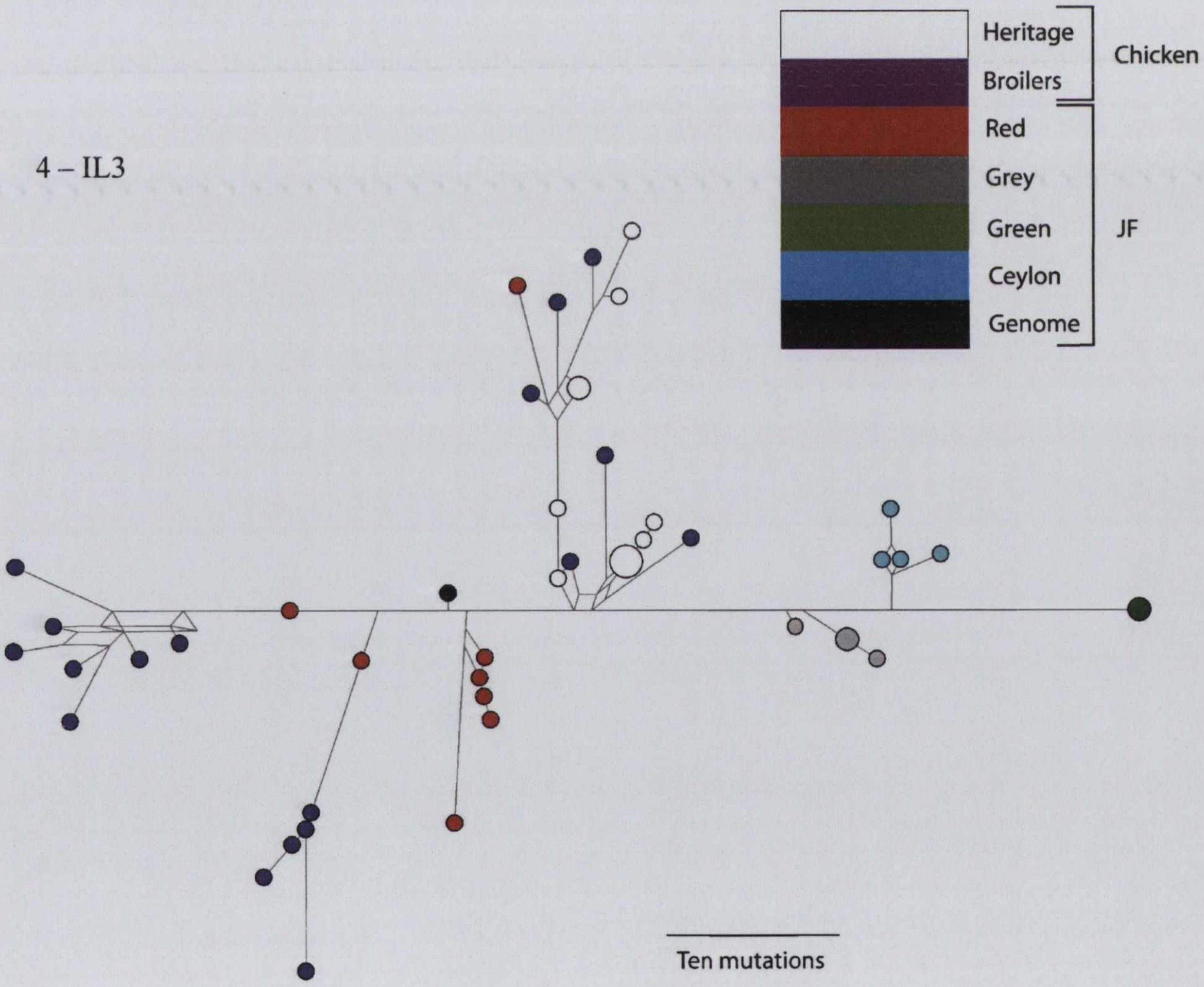
1 – GMCSF

2 – IL12A

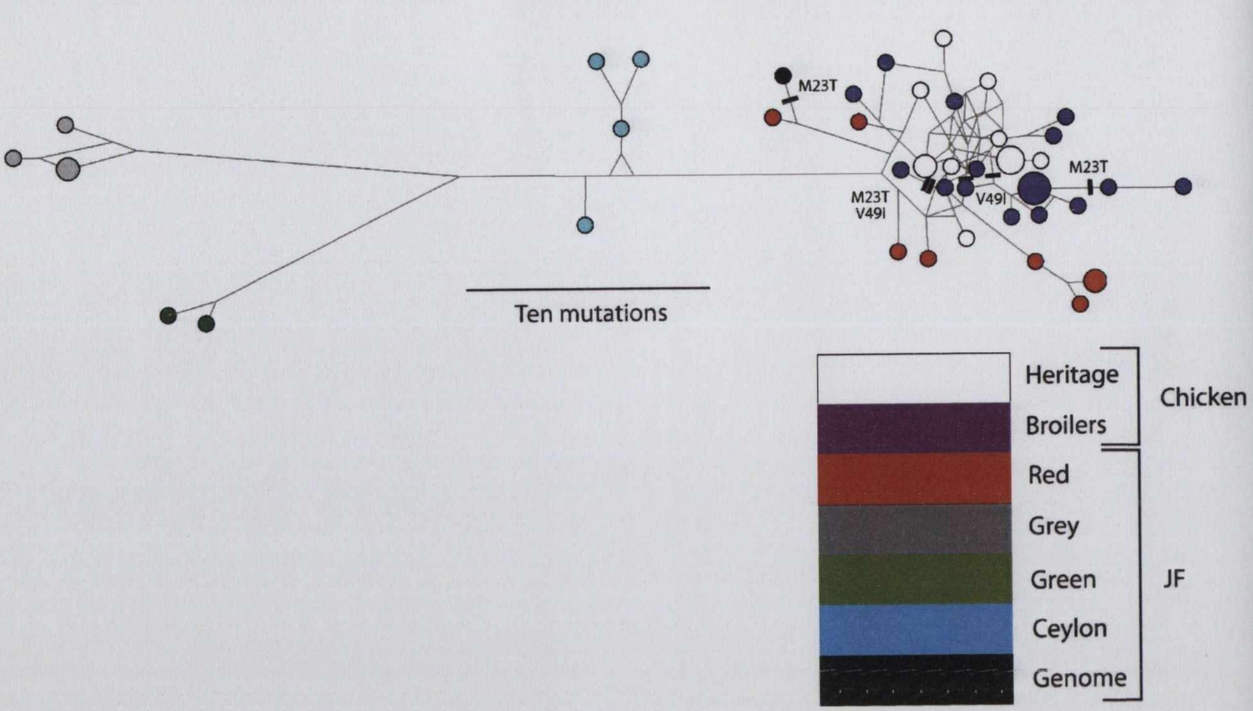
3 – IL13


A46V
A46V
A46V
Ten mutations

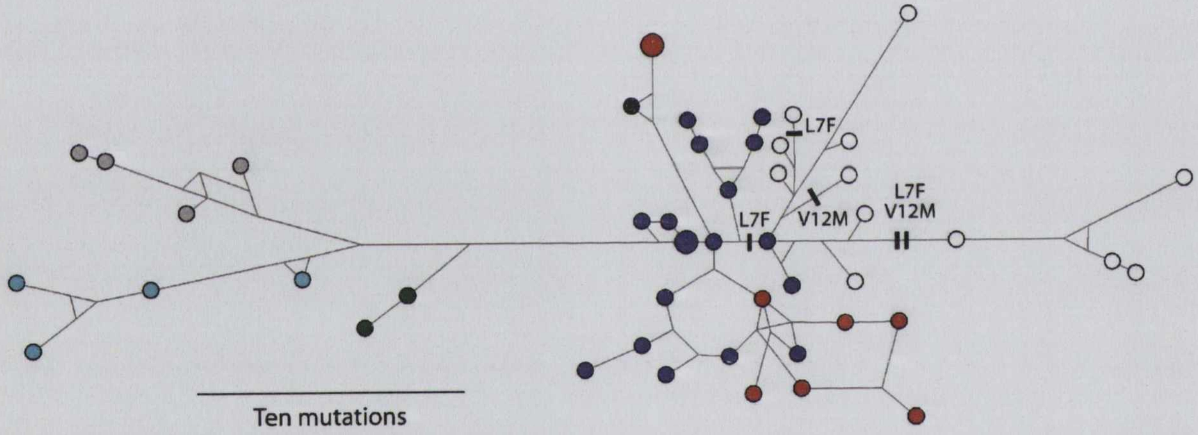
4 – IL3


Heritage
Broilers
Chicken
Red
Grey
Green
Ceylon
Genome
JF
Ten mutations

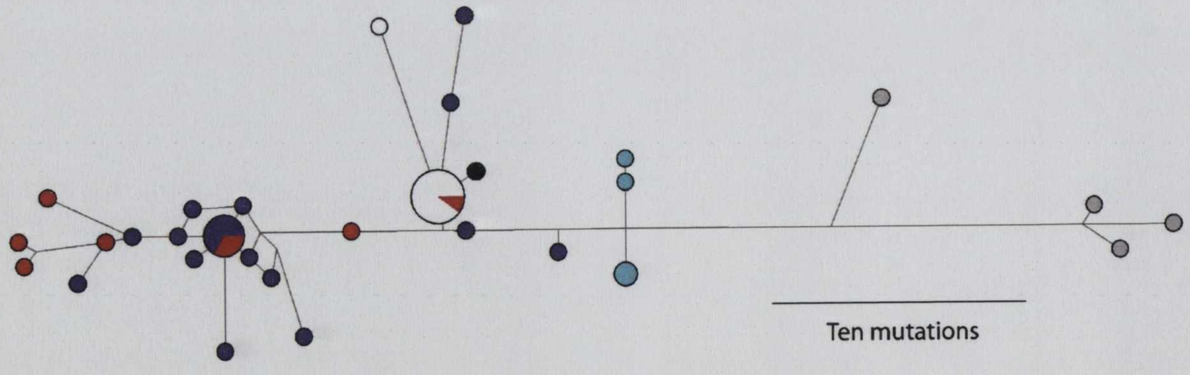
5 – IL4



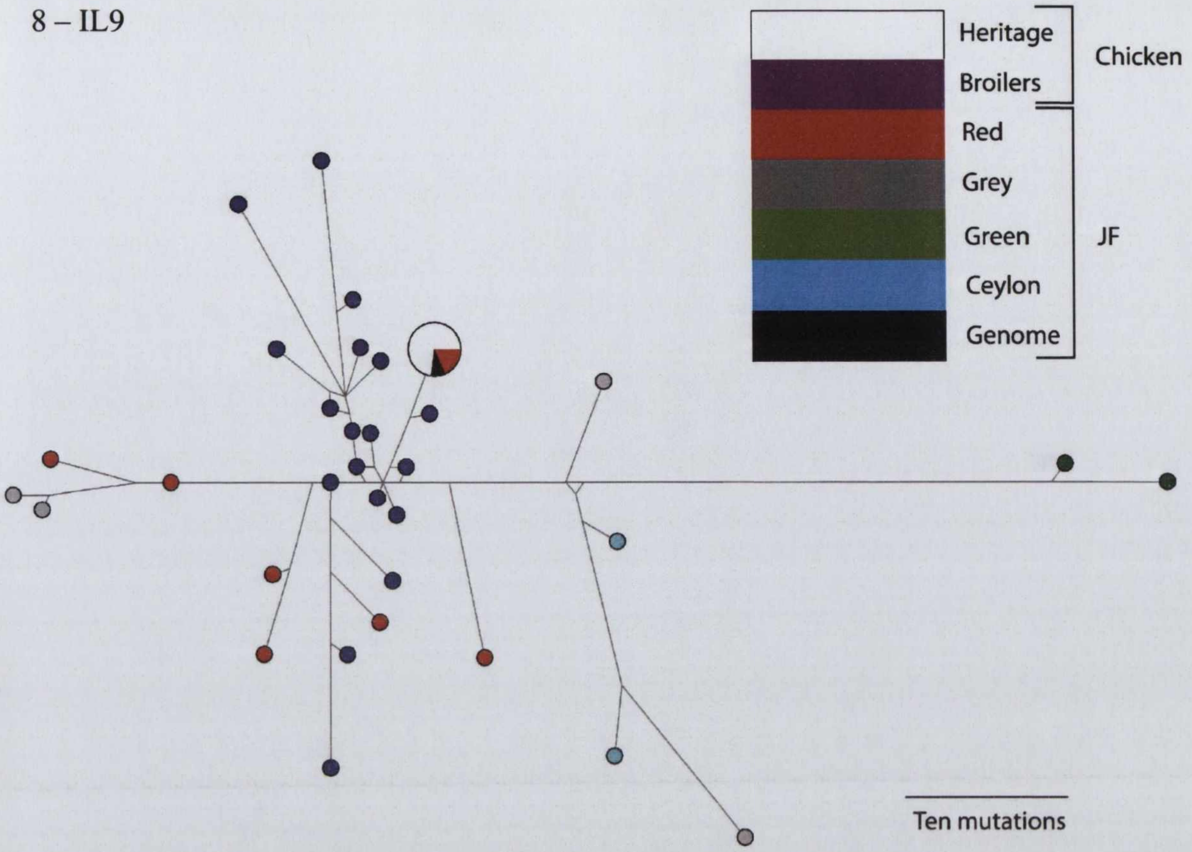
6 – IL5



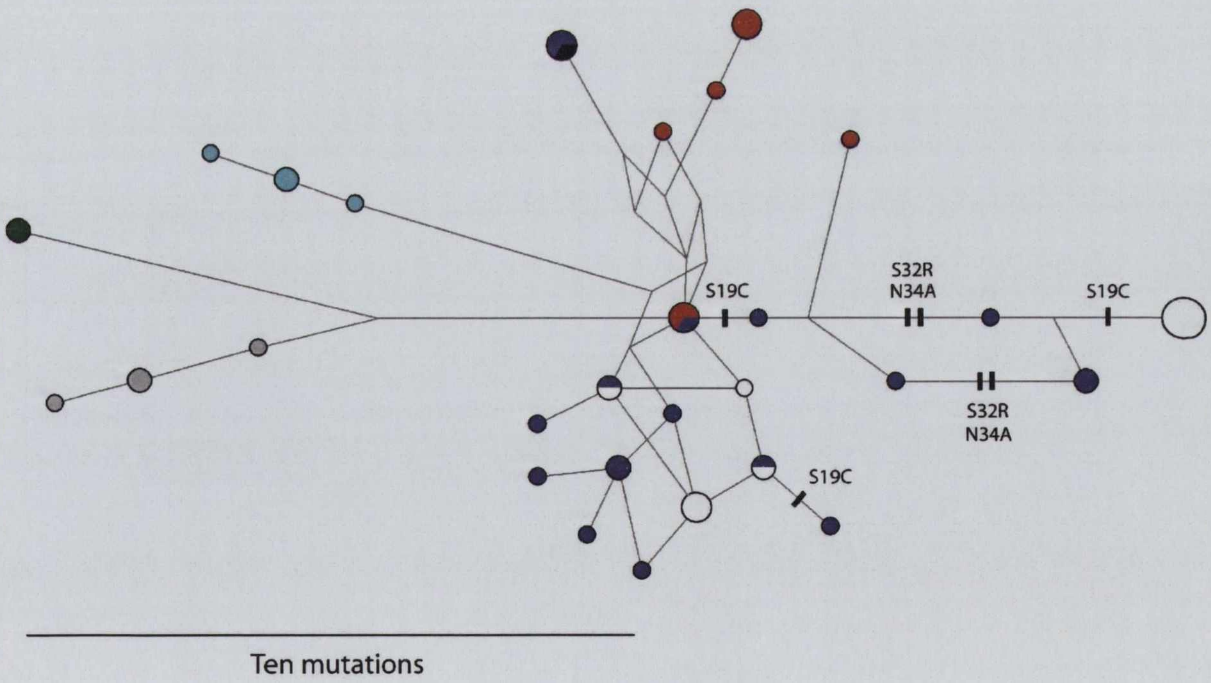
7 – IL8



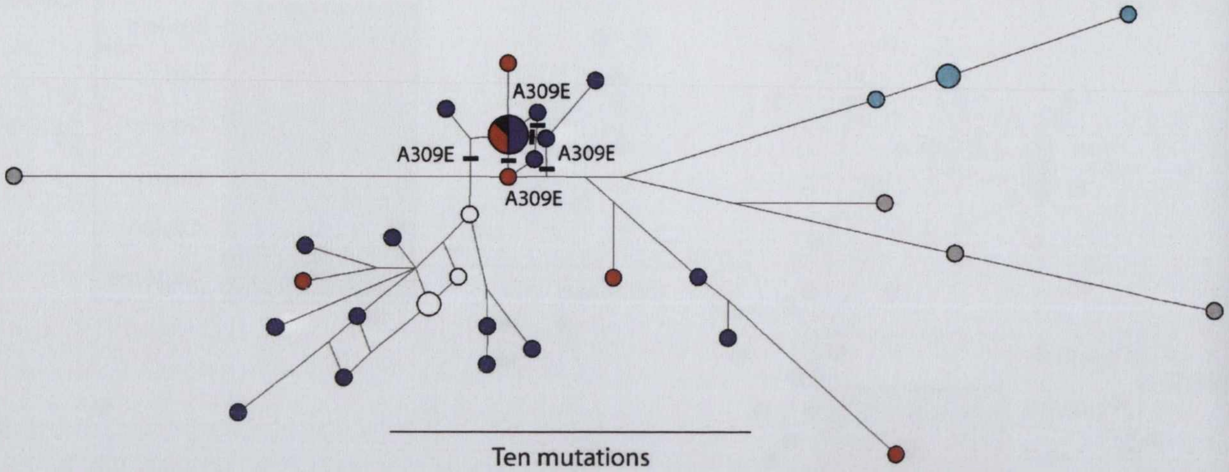
8 – IL9


Heritage
Broilers
Chicken
Red
Grey
Green
Ceylon
Genome
JF
Ten mutations

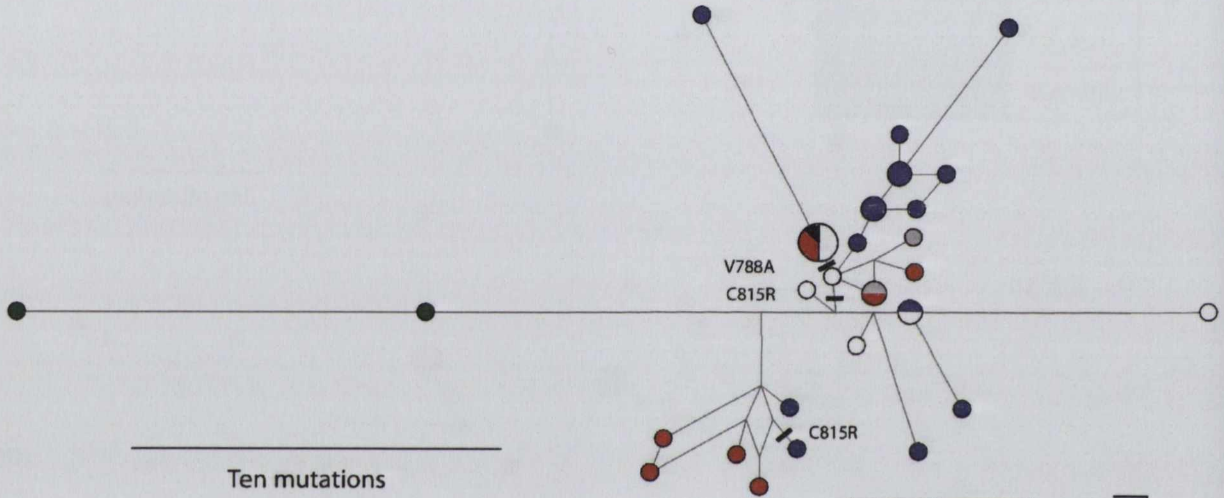
9 – KK34


S19C
S32R
N34A
S19C
S32R
N34A
S19C
Ten mutations

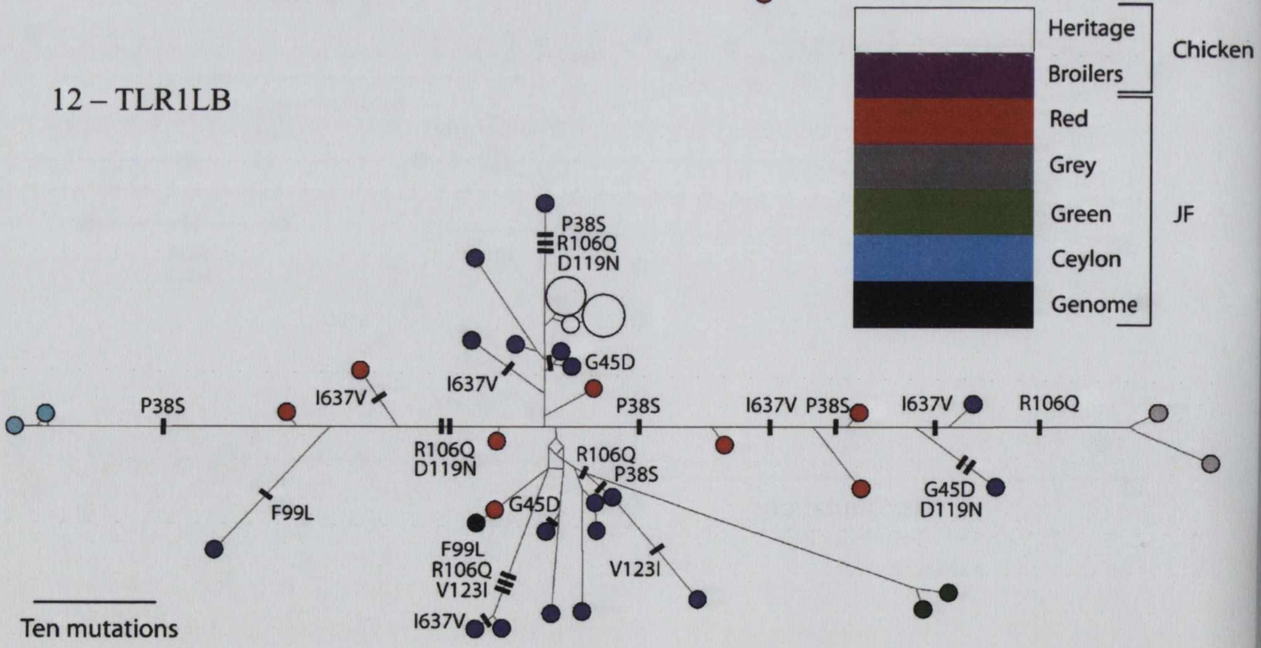
10 – TLR15



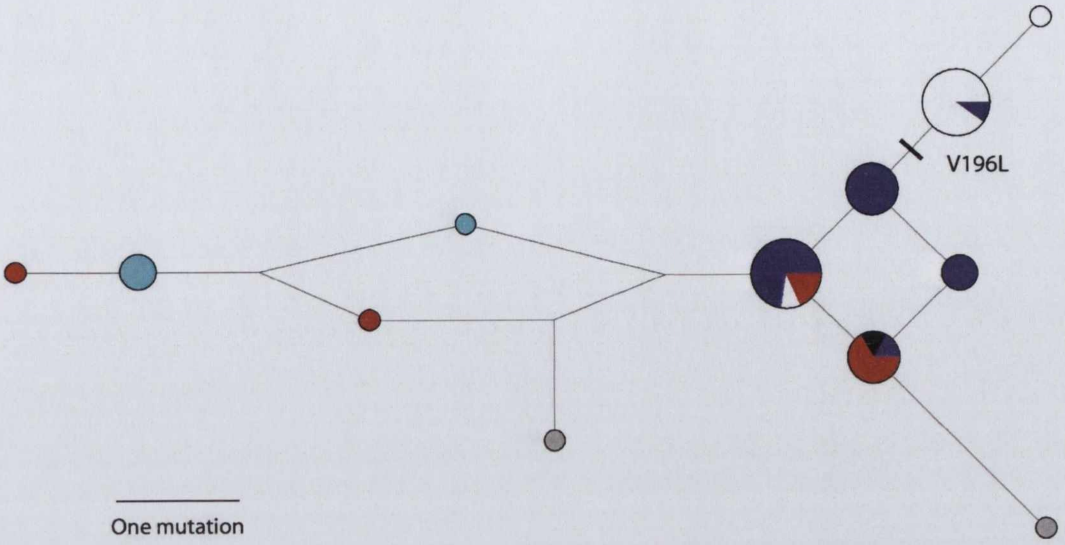
11 – TLR1LA

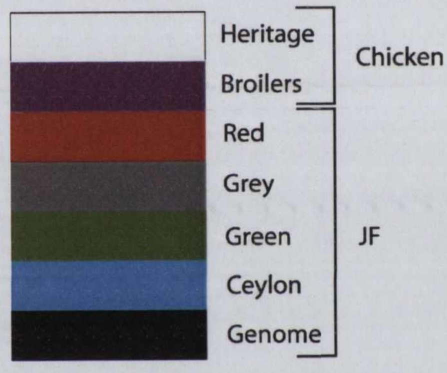


12 – TLR1LB

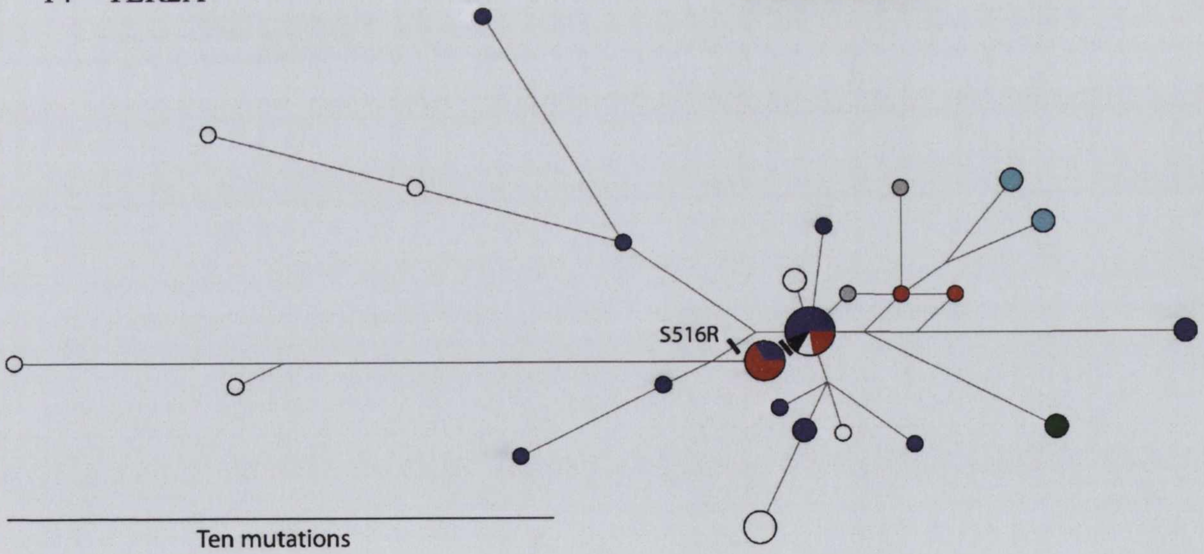


13 – TLR2B

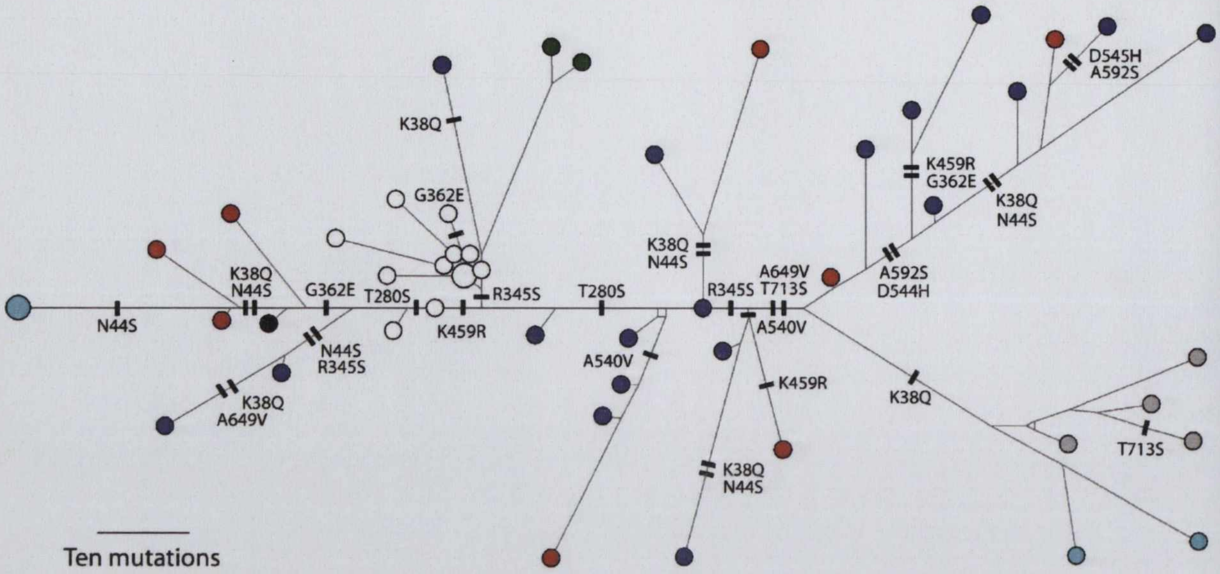




14 – TLR2A

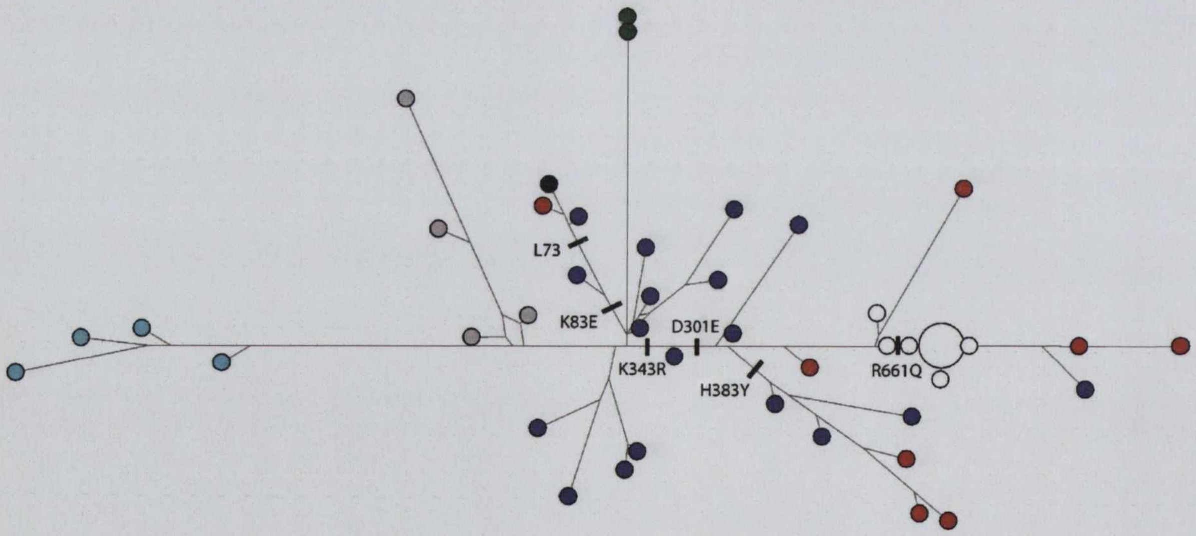


15 – TLR3

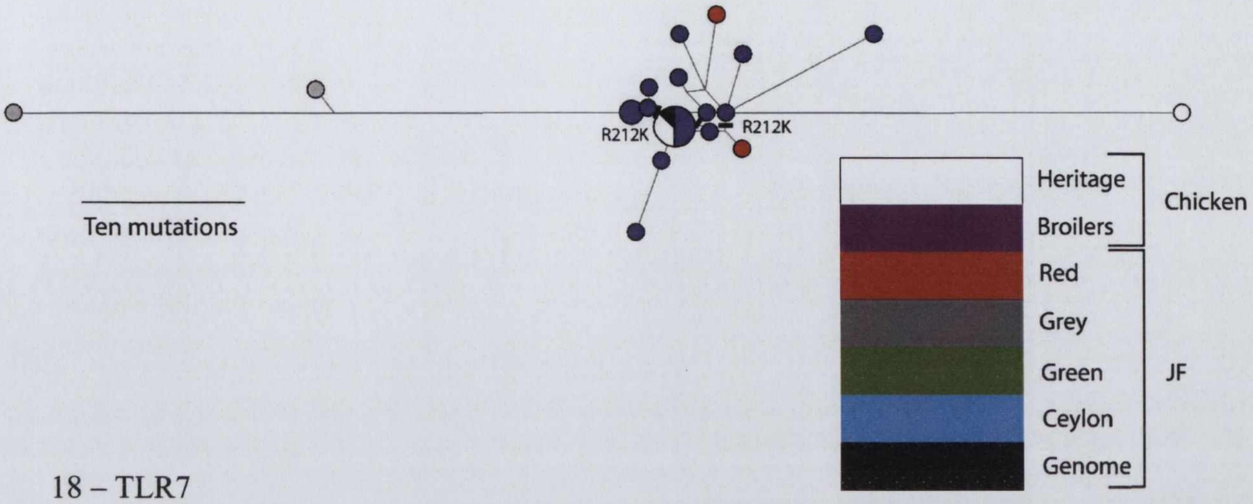


⬜	Heritage	Chicken
🟪	Broilers	
🟥	Red	JF
⬛	Grey	
🟩	Green	
🟦	Ceylon	
⬛	Genome	

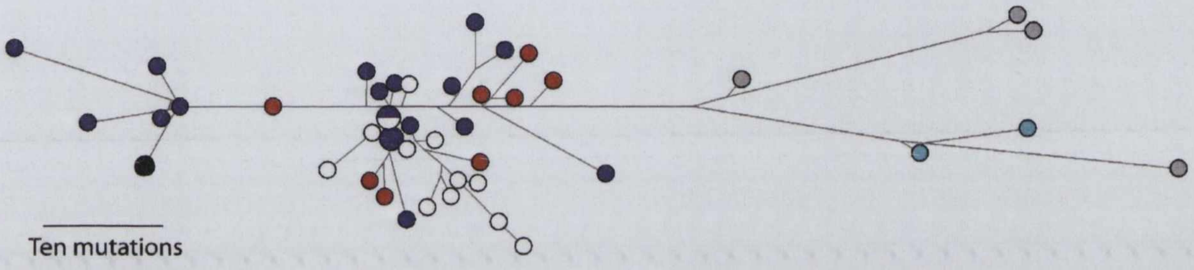
16 – TLR4



17 – TLR5



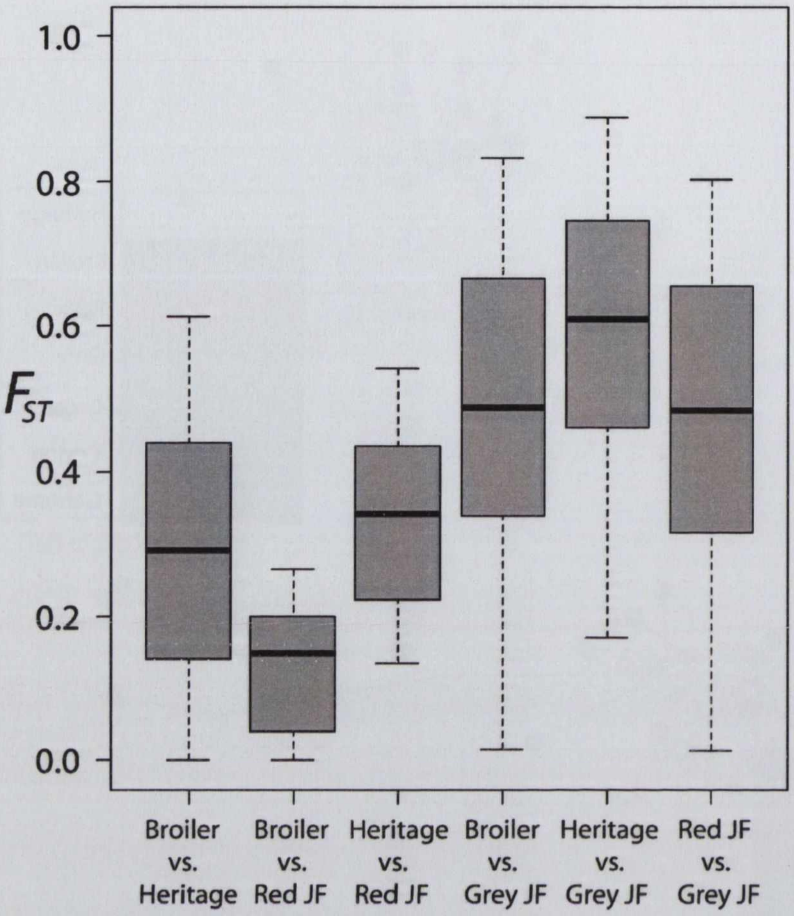
18 – TLR7



The networks showed a consistent pattern of high allele diversity. Haplotype variability within the chicken samples alone appeared to be of the same scale as that between the different taxonomical species of JF. This was consistent between gene classes and among the groups resequenced, with the exceptions of TLR2B and MC1R, which were markedly less diverse than other genes and appeared to be conserved.

Permutation tests performed with Arlequin (Schneider et al. 2000) determined whether F_{ST} values showed significant differentiation between groups. As expected, values between chicken and red JF were generally lower than those between chicken and grey JF (Figure 7.4). However, the lack of differentiation between chicken, red and grey JF initially identified in networks for TLR1LA and TLR2A was supported by F_{ST} analysis (Table 7.4). F_{ST} values between chicken and Ceylon JF as well as green JF (not shown) indicated extensive differentiation between the groups at all genes.

Figure 7.4. F_{ST} values for all genes between different sets of samples.



The F_{ST} means are indicated by the black lines. The boxes cover the area between the 1st and 3rd quartiles. The dotted lines cover the minima and maxima of the data.

7.3.2 Intraspecific chicken diversity in gene classes:

Polymorphism analysis of the sequence data identified a total of 846 SNPs in 19 genes within chicken (351 in cytokines, 492 in TLRs), 106 of which were nonsynonymous and 90 were synonymous. Tajima's *D* showed a clear difference between the cytokine (mean 0.57; Table 7.5) and the TLR (-0.62) classes (Mann-Whitney *U* $p < 0.05$; Figure 7.5; Tajima 1983b). If *D* was normally distributed, neutrality for cytokines ($p = 0.022$) and TLRs ($p = 0.025$) was rejected, but the power (β) for each was only 0.52 and 0.50, respectively. However, again if $D \sim N(0,1)$, the hypothesis that the cytokines and TLRs were in the same gene class was rejected ($p = 1.5 \times 10^{-5}$) with $\beta = 0.97$. There were no notable differences between the mean Tajima's *D* for entire genes compared to CDS regions alone (0.57 for cytokine CDS and -0.56 for TLR CDS; Table 7.6).

Table 7.4. F_{ST} values between populations and species for each gene.

Gene/Class	Broiler-H[1]	Broiler-Red[2]	H[1]-Red[2]	Broiler-Grey[3]	H-Grey[3]	Red[2]-Grey[3]
GMCSF	0.267	0.036 *	0.297	0.801	0.809	0.752
IL12A	0.404	0.143	0.435	0.546	0.614	0.535
IL13	0.225	0.152	0.253	0.645	0.692	0.668
IL3	0.324	0.266	0.421	0.547	0.664	0.631
IL4	0.171	0.243	0.310	0.831	0.851	0.802
IL5	0.189	0.200	0.183	0.682	0.655	0.655
IL8	0.613	0.000 *	0.541	0.666	0.746	0.644
IL9	0.549	0.118	0.289	0.177	0.396	0.040 *
KK34	0.096 *	0.212	0.404	0.690	0.807	0.759
TLR15	0.225 *	0.028 *	0.328	0.338	0.460	0.315
TLR1LA	0.124	0.160	0.202	0.019*	0.170*	0.013*
TLR1LB	0.440	0.043 *	0.540	0.608	0.887	0.539
TLR2B	0.487	0.171	0.531	0.015	0.313	0.275 *
TLR2A	0.058	0.041 *	0.180	0.233 *	0.257 *	0.186 *
TLR3	0.355	0.040 *	0.390	0.450	0.678	0.496
TLR4	0.550	0.052	0.451	0.371	0.736	0.393
TLR5	0.000 *	0.178	0.136 *	0.591	0.512	0.597
TLR7	0.141	0.005 *	0.223	0.529	0.619 *	0.536 *
Cytokine	0.315	0.152	0.348	0.621	0.693	0.610
TLR	0.266	0.080	0.331	0.367	0.506	0.389
All 18	0.290	0.116	0.339	0.494	0.599	0.499

[1] Heritage chickens. [2] Red JF. [3] Grey JF. * denotes where p values were not significant, indicating no significant differentiation between the two populations examined. Values for Ceylon JF and green JF are not listed as there was lower sampling for these subspecies, no haplotype network evidence of introgression, and their F_{ST} values with other groups were significantly high.

Table 7.5. Mean diversity and summary statistic values for the gene classes.

Group	π[1]	θ_W[2]	K[2]	π/K	π_A/π_S	K_A/K_S	$\frac{\pi_A/\pi_S}{K_A/K_S}$	Tajima's D
Cytokine	3.44 ± 2.16	3.20 ± 2.55	5.681	0.611	0.961	0.257	4.264	0.572 ± 0.88
TLR	2.13 ± 1.12	2.73 ± 1.50	3.398	0.649	0.305	0.375	0.875	-0.619 ± 0.95
Cytokine & TLR	2.79 ± 2.61	2.97 ± 2.98	4.709	0.610	0.574	0.316	2.434	-0.039 ± 1.02

[1] π, θ_W and K are measured per kb. The average number of SNPs observed was 39.0 at cytokine genes, 54.7 at TLRs – a mean of 46.8 between classes.

Correlations using Pearson's correlation coefficient (r) between Tajima's D and resequenced gene length ($r^2 = 0.09$ for cytokines, $r^2 = 0.08$ for TLRs) or with CDS length ($r^2 = 0.01$ for cytokines, $r^2 = 0.06$ for TLRs) showed no significant association, indicating that although relatively more TLR gene and coding region was resequenced, this did not significantly bias D. Although Mann-Whitney U-tests found that divergence between species, measured with *K*, was lower for TLRs (5.7 per kb vs 3.4, $p < 0.05$), π/K was almost the same for each class (0.61 vs 0.65; Table 7.7), meaning relative background mutation rates and JF were the same between classes.

Table 7.6. Selected descriptive and summary statistics, and tests of neutrality in chicken at each gene.

Test/	H[1]	π	θ_W	*Hd*	Tajima's *D*		Fu's F_S	Fay &	Fu & Li's	
Gene					Gene[2]	CDS[3]		Wu's *H*	*D*	*F*
GMCSF	15	1.93	1.39	0.940	1.299	1.319	-3.091	-0.543	0.794	1.158
P value	*ns*	*ns*	*ns*	*ns*	*0.017*	*ns*	*ns*	*ns*	*ns*	*ns*
IL12A	20	4.45	4.15	0.894	0.356	-1.550	-4.758	-2.722	-0.400	-0.204
P value	*ns*	*ns*	*ns*	*0.001*	*ns*	*0.012*	*ns*	*ns*	*ns*	*ns*
IL13	25	3.98	2.85	0.989	1.454	1.862	-11.836	-3.307	1.720	1.979
P value	*0.019*	*ns*	*0.054*	*ns*	*<0.001*	*0.010*	*0.074*	*ns*	*<0.001*	*<0.001*
IL3	26	2.67	1.93	0.984	1.427	-	-8.906	5.205	1.267	1.616
P value	*<0.001*	*ns*	*ns*	*ns*	*<0.001*	-	*ns*	*ns*	*0.015*	*ns*
IL4	24	3.12	2.15	0.977	1.551	0.869	-14.597	0.542	1.738	2.020
P value	*ns*	*ns*	*ns*	*ns*	*0.001*	*ns*	*ns*	*ns*	*0.001*	*<0.001*
IL5	29	2.42	2.71	0.998	-0.389	-0.053	-25.101	1.370	0.136	-0.061
P value	*0.015*	*ns*	*ns*	*ns*	*ns*	*ns*	*0.045*	*ns*	*ns*	*ns*
IL8	17	2.31	2.64	0.889	-0.462	-	-3.645	-1.915	-1.630	-1.486
P value	*ns*	*ns*	*ns*	*ns*	*ns*	-	*ns*	*ns*	*0.044*	*ns*
IL9	19	8.93	10.00	0.992	-0.460	1.319	-2.014	-2.646	-0.863	-1.068
P value	*0.011*	*ns*	*ns*	*<0.001*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
KK34	18	1.13	1.02	0.972	0.370	0.255	-8.423	1.526	1.349	1.238
P value	*0.033*	*ns*	*ns*	*0.012*	*ns*	*ns*	*0.038*	*ns*	*0.031*	*ns*
TLR15	20	1.89	1.86	0.991	0.062	-0.315	-14.004	1.524	0.240	0.221
P value	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
TLR1LA	20	1.68	3.02	0.972	-1.679	-1.656	-10.080	-1.797	-3.210	-3.251
P value	*0.004*	*ns*	*0.010*	*0.022*	*0.002*	*0.010*	*0.013*	*ns*	*<0.001*	*<0.001*
TLR1LB	21	3.20	3.34	0.943	-0.159	-0.217	-1.365	-23.126	1.444	1.036
P value	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*0.014*	*0.010*	*ns*
TLR2B	6	0.33	0.27	0.772	0.601	0.776	-0.996	0.009	0.010	0.217
P value	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
TLR2A	18	0.70	1.13	0.943	-1.405	-1.123	-4.686	2.345	-0.859	-1.287
P value	*0.003*	*ns*	*0.031*	*ns*	*0.003*	*ns*	*ns*	*ns*	*ns*	*ns*
TLR3	29	4.17	4.68	0.998	-0.423	-0.546	-9.180	3.292	-1.551	-1.385
P value	*0.001*	*ns*	*ns*	*0.016*	*ns*	*ns*	*ns*	*0.019*	*0.007*	*0.018*
TLR4	24	2.61	2.35	0.952	0.422	0.567	-5.101	-0.322	0.832	0.834
P value	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
TLR5	14	2.62	5.28	0.916	-2.040	-2.009	-2.555	6.189	-4.184	-4.239
P value	*0.019*	*ns*	*0.027*	*ns*	*ns*	*0.008*	*0.008*	*0.011*	*<0.001*	*<0.001*
TLR7	28	1.99	2.66	0.995	-0.950	-	-14.110	-7.945	-1.386	-1.485
P value	*<0.001*	*ns*	*ns*	*0.001*	*ns*	-	*<0.001*	*ns*	*ns*	*0.044*

[1] Number of haplotypes. [2] Values for the entire resequenced region. [3] Values for the resequenced coding region only; CDS *D* values could not be determined for genes with no coding SNPs (IL3, IL8 and TLR7). P values were determined by coalescent simulation using DnaSP: statistics for which $p > 0.05$ are shown as "ns". π, θ_W and *K* are measured per kb. Coalescent simulations for TLR2B, where recombination was too high to be accurately estimated, used $R = 500$, the maximum allowed (see Table 7.9 for values).

Figure 7.5 Tajima's *D* for the cytokine and TLR classes.

The thick black lines indicate the means. The boxes cover the area between the 1st and 3rd quartiles. The dotted lines cover the minima and maxima of the data

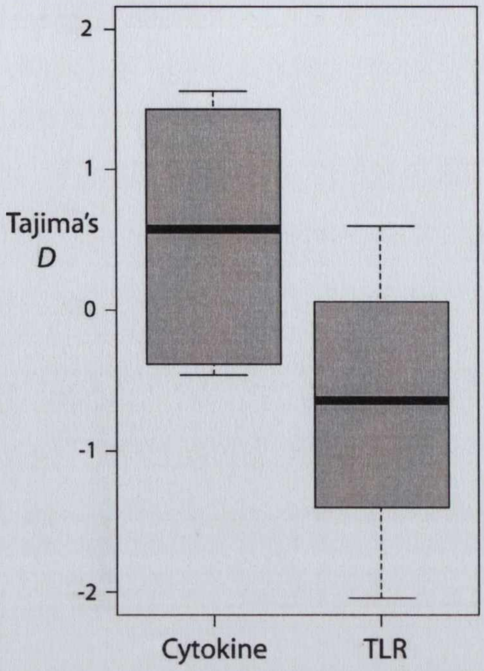


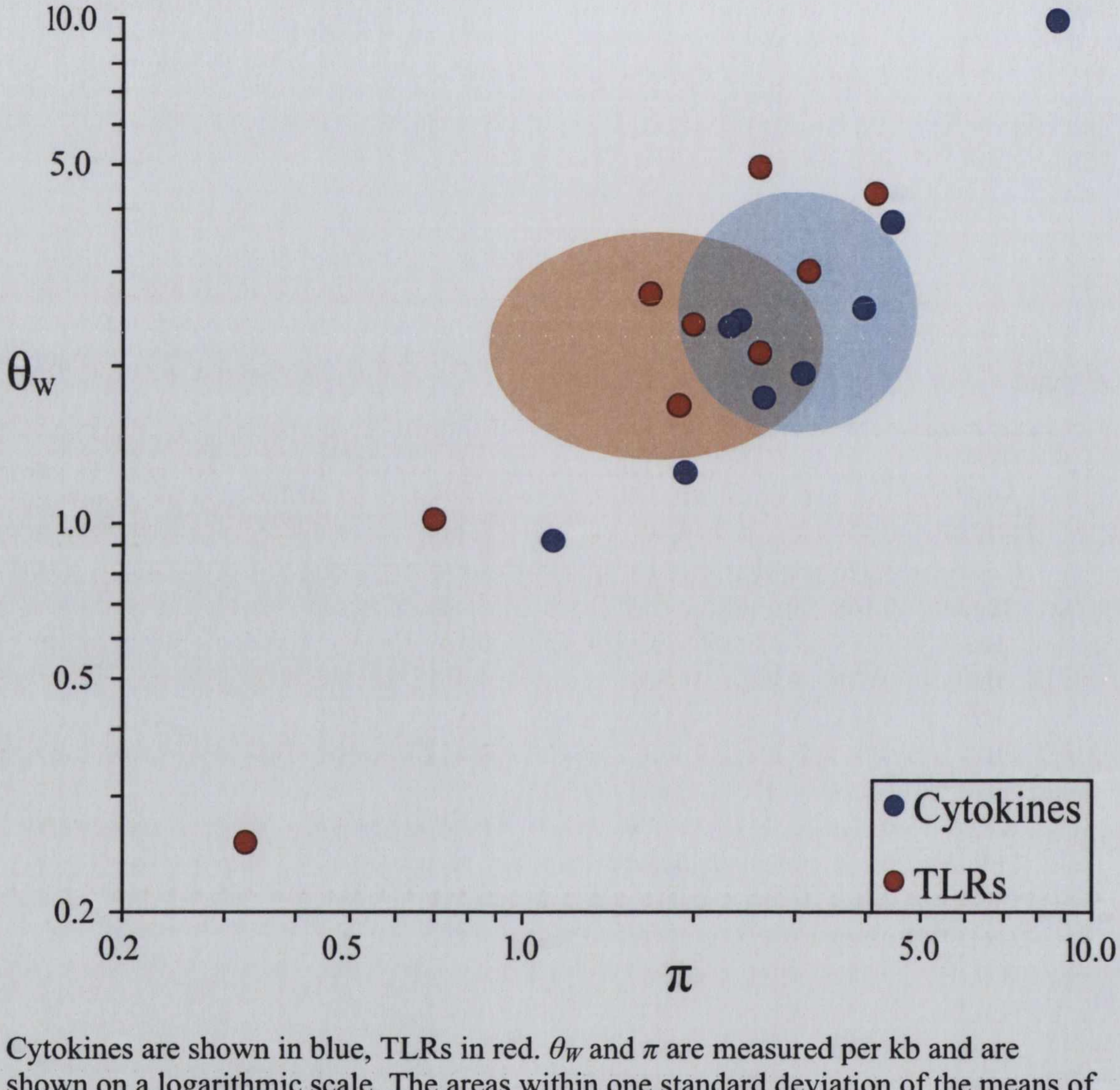
Table 7.7. SNP frequency, polymorphism and divergence rates in chicken at each gene.

Test/ Gene	π_A/π_S	K_A/K_S	$\frac{\pi_A/\pi_S}{K_A/K_S}$	K	π/K	nonsyn [1]	syn [2]	aA [3]	S [4]
						SNPs			
GMCSF	0.338	0.341	0.991	4.42	0.437	1	1	554	15
IL12A	3.276	0.176	18.61	9.84	0.452	5	1	202	33
IL13	0.108	0.166	0.651	5.99	0.664	1	3	138	31
IL3	0	0	0	3.55	0.752	0	0	138	40
IL4	0.636	0.835	0.762	9.13	0.342	2	1	221	72
IL5	3.462	0.250	13.85	4.61	0.525	2	1	61	54
IL8	0	0	0	4.33	0.533	0	0	103	32
IL9	0.433	0.178	2.433	7.27	1.228	2	5	138	57
KK34	0.396	0.367	1.079	1.99	0.568	5	5	169	17
TLR15	0.160	0.230	0.696	3.35	0.564	8	10	868	21
TLR1LA	0.540	0.403	1.340	2.30	0.730	18	5	245	41
TLR1LB	0.629	0.478	1.316	3.62	0.884	11	3	263	79
TLR2B	0.347	0.256	1.355	1.04	0.317	2	1	732	4
TLR2A	0.230	0.529	0.435	0.67	1.045	8	8	793	40
TLR3	0.350	0.424	0.825	5.88	0.709	25	20	840	122
TLR4	0.160	0.181	0.884	4.42	0.590	9	12	843	58
TLR5	0.326	0.318	1.025	5.43	0.483	22	17	861	59
TLR7	0	0.885	0	3.87	0.514	0	0	105	68

[1] Number of nonsynonymous SNPs. [2] Number of synonymous SNPs. [3] Number of amino acids. [4] Total number of SNPs. Genes with no nsSNPs (IL3, IL8 and TLR7) also had no observed synonymous SNPs (sSNPs). Although MC1R was included in the total SNP counts, only three noncoding and no coding SNPs were observed within chicken.

When the linear relationship between π and θ_W was examined, a trend of a higher π than θ_W for cytokines and a higher θ_W than π for TLRs was observed (Figure 7.6). The apparent cause of the more positive Tajima's *D* in cytokines was their higher average π (3.44 per kb vs 2.13 for TLRs; Figure 7.7) – θ_W was more similar than π between the gene classes (3.20 per kb vs 2.73).

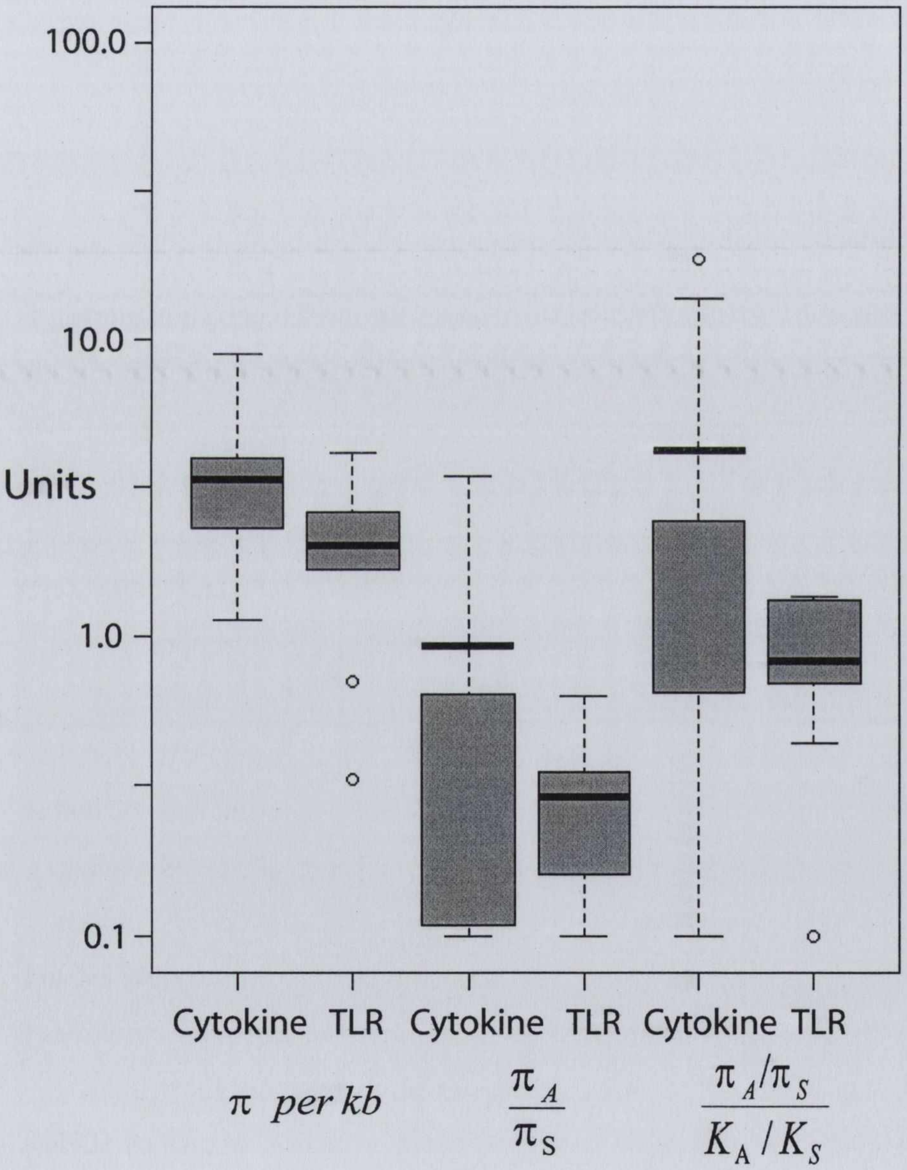
Figure 7.6. θ_W versus π for cytokine and TLR genes.



Cytokines are shown in blue, TLRs in red. θ_W and π are measured per kb and are shown on a logarithmic scale. The areas within one standard deviation of the means of each class are shown: pale blue for cytokines and pink for TLRs.

Tajima's *D* values were calculated for the predicted TLR transmembrane (-1.19), cytoplasmic (-0.75) and extracellular (-0.29) domains. These indicated stronger directional selection on the gene encoding the protein portion spanning the cell membrane. However, there were no polymorphisms at TLR1LA, TLR1LB, TLR2B and TLR7 in the transmembrane domain; at TLR2A and TLR7 in the cytoplasmic domain and at TLR7 in the extracellular domains. When this sampling issue is coupled with the shortness of the transmembrane regions, it is clear that additional genes would need to be resequenced to examine the evolutionary patterns within the TLR domains in a robust manner. Similarly, IL3, IL8 and IL9 had no mutations in their predicted signal peptides, so evaluating *D* for the short signal and active protein regions did not have sufficient variable sites to be informative.

Figure 7.7. π per kb, π_A/π_S and $\pi_A/\pi_S/K_A/K_S$ for the cytokine and TLR classes.



Note that values are plotted on a log scale. The thick black lines indicate the means. The boxes cover the area between the 1st and 3rd quartiles. The dotted lines cover the minima and maxima of the data.

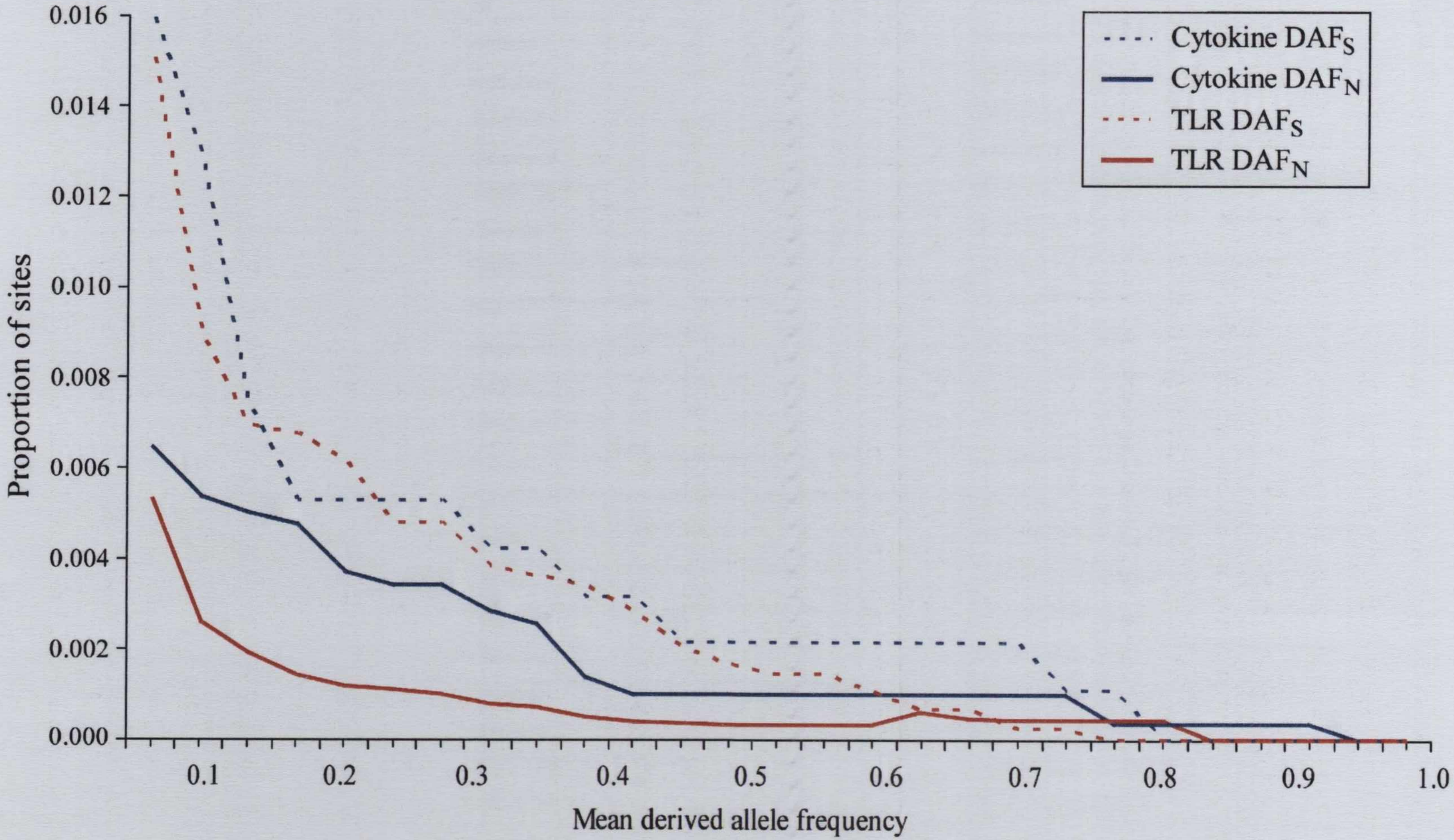
7.3.3 Functional SNPs and the allele frequency spectrum:

In order to estimate the fraction of segregating nonsynonymous SNPs that might be deleterious, DAF_N/DAF_S was determined for chicken alleles in each gene class segregating at low (< 0.2) and high (≥ 0.2) frequencies (Liti et al. 2009). DAF_N/DAF_S < 1 for both classes, indicating greater purifying selection at nonsynonymous sites. DAF_N/DAF_S for high compared to that for low frequencies gives an estimate of the fraction of segregating deleterious alleles: so for cytokines (low 0.440 minus high 0.226 = 0.214), it was less than that for TLRs (low 0.556 minus high 0.183 = 0.373). This suggested that relatively more of the nonsynonymous SNPs present at cytokines may be of functional relevance. The average estimated fraction of replacement neutral substitutions was about the same for cytokines ($f = 0.40 \pm 0.09$) and TLRs (0.36 ± 0.12; Table 7.9; Smith & Eyre-Walker 2002).

The allele frequency spectra of DAF_N and DAF_S in chickens showed that the proportion of sites segregating decreased with increasing allele frequency in both gene classes, as expected (Figure 7.8), despite the significant abundance of alleles, as indicated by the negative Fu's F_S values and very high *Hd* values (Table 7.6). For $0.1 \leq DAF \leq 0.4$, the rate at synonymous sites was similar between classes, whereas that at nonsynonymous sites was lower for TLRs compared to cytokines. For $DAF > 0.4$, although DAF_N was also higher for cytokines, DAF_S was also higher, indicating that the difference in DAF_N values was likely to be neutral. When DAF_N and DAF_S were adjusted for the proportion of sites segregating in all chicken genotypes, the difference between the two classes was clearer (Figure 7.9), and was illustrated further by using π_A/π_S instead of π_A and π_S (Figure 7.10). These results indicated that while variation at synonymous sites did not differ noticeably between classes, that at nonsynonymous sites was higher at intermediate frequencies for the cytokine class.

The functional relevance of the difference between Tajima's *D* for the cytokines and TLRs was illustrated by the correlations between DAF_N per kb and *D* for cytokines ($r^2 = 0.44$; Figure 7.11) and TLRs ($r^2 = 0.35$). A greater number of nonsynonymous variants was linked to an excess of low- compared to intermediate-frequency alleles, illustrating that part of the negative Tajima's *D* at TLRs is associated with the number of deleterious alleles present.

Figure 7.8. Derived allele frequency spectrum as K_A and K_S in the chicken samples.



The derived allele frequency spectrum is shown as the proportion of SNPs segregating at nonsynonymous (K_A) and synonymous (K_S) sites for each gene class.

Figure 7.9. Population-wide decay of segregating variable sites as π_A and π_S at a given chicken DAF for the cytokine and TLR gene classes.

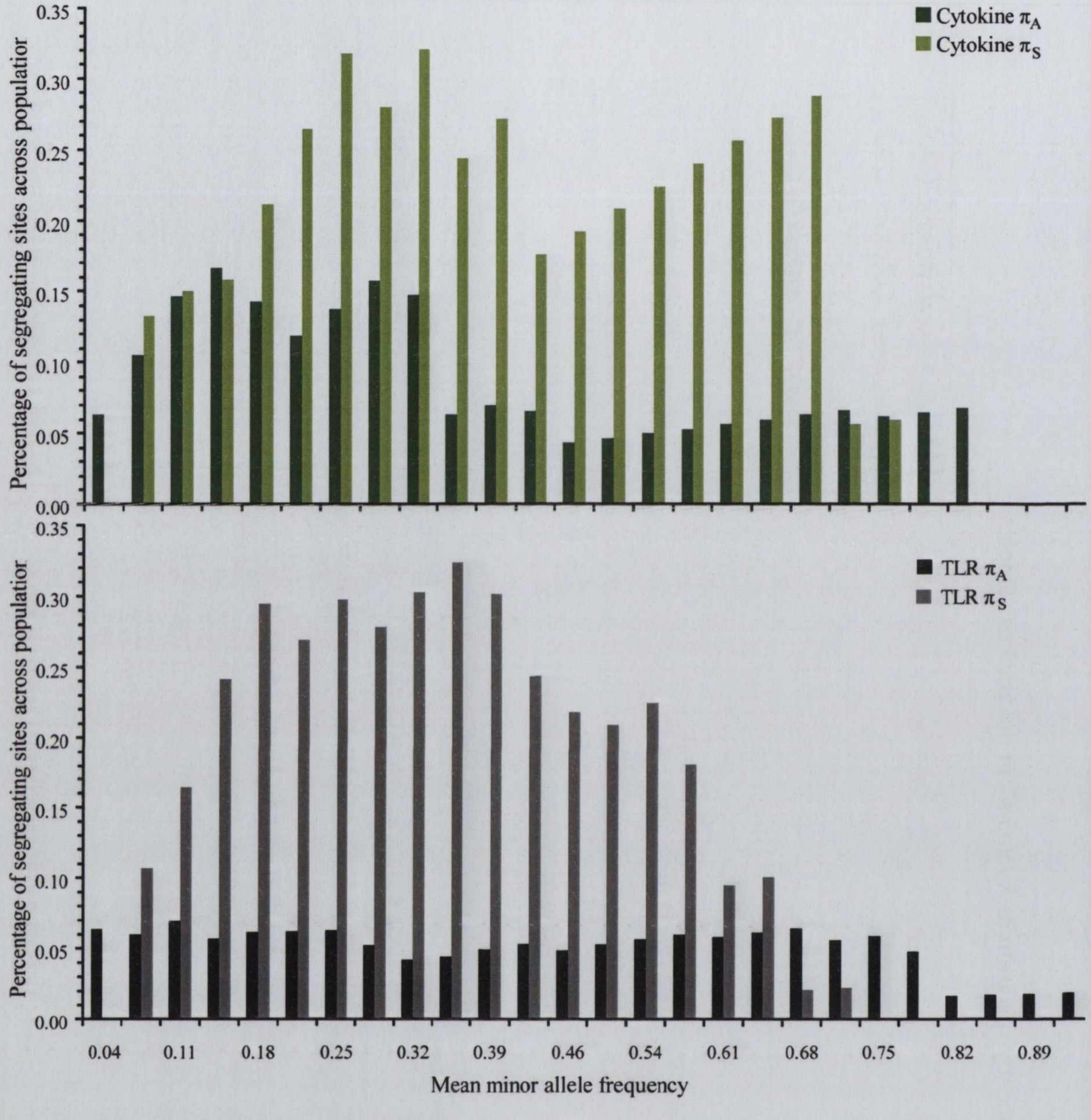
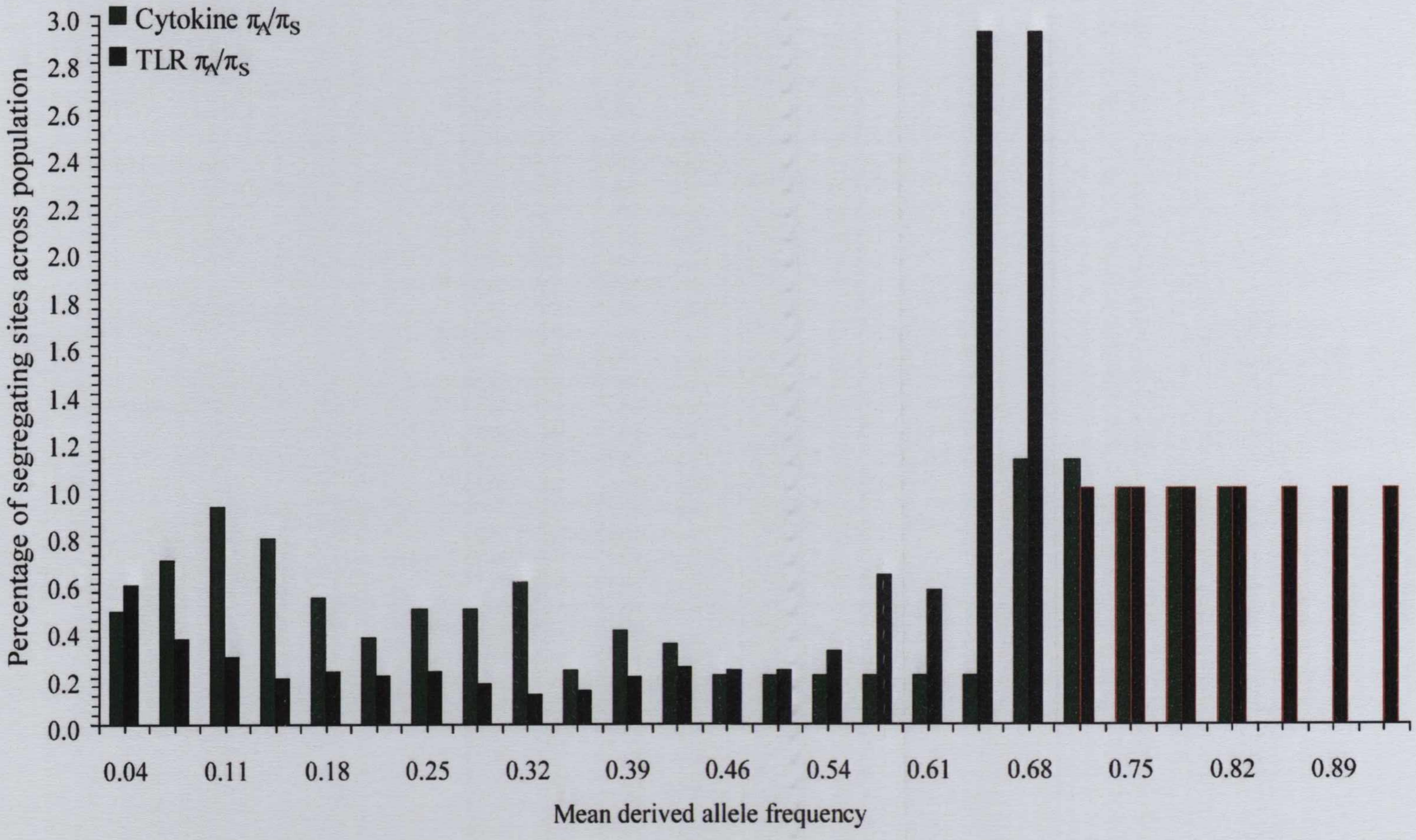
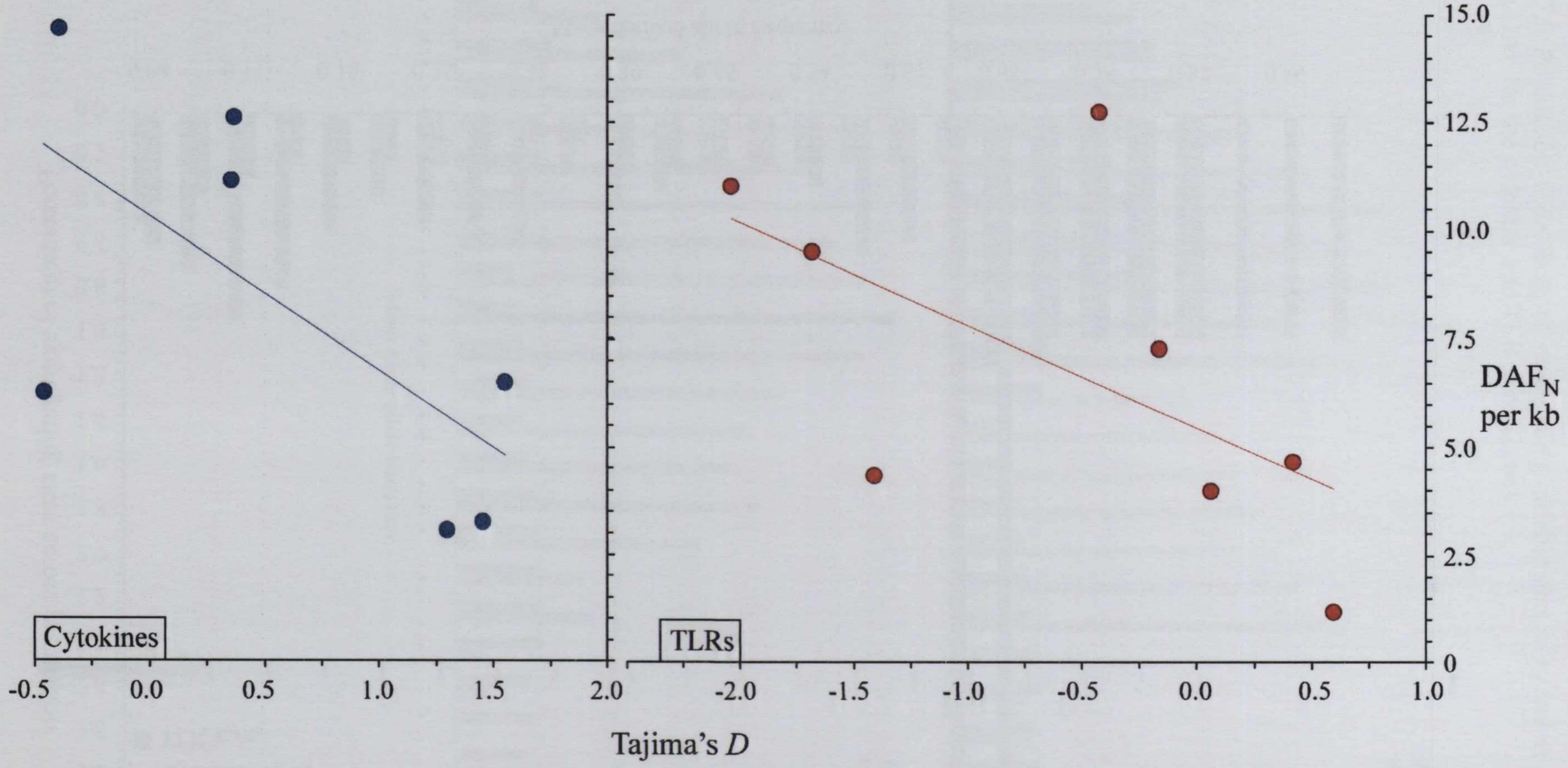


Figure 7.10. Population-wide segregating variable sites as π_A/π_S at chicken DAFs.



Where DAF > 0.73 for cytokines (green) and DAF > 0.70 for TLRs (black), values have $\pi_S = 0$ and $\pi_A > 0$ and are shown with red outlines to be fixed as $\pi_A/\pi_S = 1$.

Figure 7.11. Tajima's D vs $\mathrm{DAF_N}$ in the chicken samples for each gene.



Cytokine genes (left) are in blue and TLR genes (right) are in red. The linear trend lines for each gene class indicate D and $\mathrm{DAF_N}$ correlate for both (cytokines, $r^2 = 0.437$; TLRs, $r^2 = 0.353$). Genes with $\mathrm{DAF_N} = 0$ were not included (IL3, IL8, TLR7).

7.3.4 The effect of demographic change on variation:

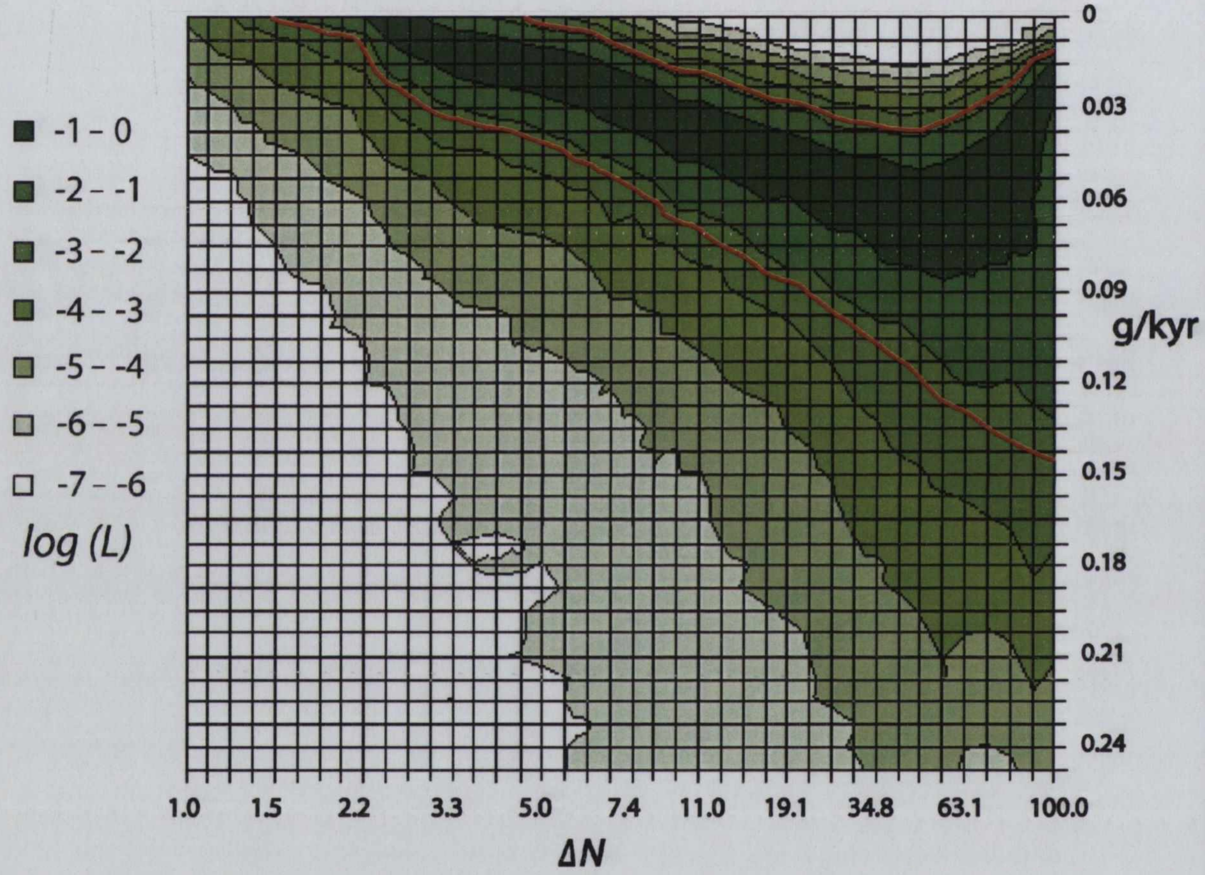
In order to test if the chicken's complex population history could account for the observed patterns of diversity, the cytokine and TLR genes' observed data was simulated using MS (Hudson 2002). Following an ancestral domestication event $0.1N_0$ kya, these simulations varied the rates of chicken population size growth (ΔN) and introgression of red JF genotypes into the chicken population (g/kyr) – N_0 is the ancestral population size prior to domestication. The likelihoods of the models were calculated in the context of three Tajima's D values: neutral ($D = 0$), cytokine ($D = 0.572$) and TLR ($D = -0.612$). LRTs were conducted for $\chi^2 = 2(L_1-L_0)$ with one degree of freedom where L_1 was the model with the maximum likelihood and L_0 was an alternative model. These values were used to determine the ranges of introgression of red JF into the chicken population (g/kyr) and the change in the chicken population size after domestication (ΔN) for which the optimal model was significantly better ($p < 0.05$).

For the neutral D, the most positive log likelihood ($log_e(L)$) at -0.248 had an introgression rate of 0.06 of genotypes per kyr and $\Delta N = 19.6$, so models with $log_e(L) < -2.168$ had $p < 0.05$ and thus were not significantly different from the model with the maximum likelihood (Figure 7.12). For the cytokine D, an introgression rate of 0 g/kyr and $\Delta N = 16.4$ had the maximum likelihood of -0.089, so models with $log_e(L) < -1.991$ had $p < 0.05$ (Figure 7.13). And for the TLR D, introgression at 0.165 g/kyr and $\Delta N = 1$ had the peak likelihood of -0.802, so models with $log_e(L) < -2.723$ had $p < 0.05$ (Figure 7.14). The demographic history simulated here could not account for the Tajima's D values observed for the cytokine and TLR classes ($p = 0.014$; Figure 7.15) and the overlap between the areas significantly more likely for the neutral and cytokine ($p = 0.062$) or neutral and TLR ($p = 0.057$) was minimal.

For the most likely parameters for each gene class, the simulated Tajima's D values for each (neutral = 0.01; cytokine = 0.50; TLR = -0.32) were close to the observed values, though the simulations could not entirely account for the observed TLRs' Tajima's D (-0.62). Even for their optimal Tajima's D value model, the cytokine gene class had a lower number of SNPs observed than simulated (34 vs 30; Table 7.8) and TLRs had more (51 vs 54). The observed cytokine π (3.44 per kb) was still much higher than that for the simulated neutral (2.41) or cytokine D values (2.74). For the

TLRs, observed π was lower (2.13) than the neutral (2.54) or TLR (2.32) values. These simulations provided evidence that complex demographic history may not alone account for the unusual evolution of these two gene classes, however, it did not refute the idea that either class could be selectively neutral.

Figure 7.12. Likelihood contour map of a neutral Tajima's *D* value (0) for introgression of red JF genotypes into the chicken population (g/kyr) and chicken population size increase (ΔN), models with $log_e(L) < -1.991$ were significantly less likely.



G/kyr is the fraction of red JF genotypes migrating into the chicken population per kyr. ΔN is the chicken population size increase since domestication such that $\Delta N = e^{-\alpha t}$. The contour lines denote changes in the likelihood of the models simulated according to ΔN and g/kyr, where lighter areas were less likely. The maximum likelihood model was at g/kyr = *0.060* and $\Delta N = 19.6$, where the log likelihood ($log_e(L)$) = *-0.248*, so according to LRTs, models with $log_e(L) < -2.168$ were significantly less likely ($p < 0.05$) – these are denoted by red lines.

Figure 7.13. Contour map of the simulated cytokine model (Tajima's $D = 0.572$) log likelihood ($log_e(L)$) for introgression of red JF genotypes into the chicken population (g/kyr %) and chicken population size increase (ΔN), models with $log_e(L) < -1.991$ were significantly less likely.

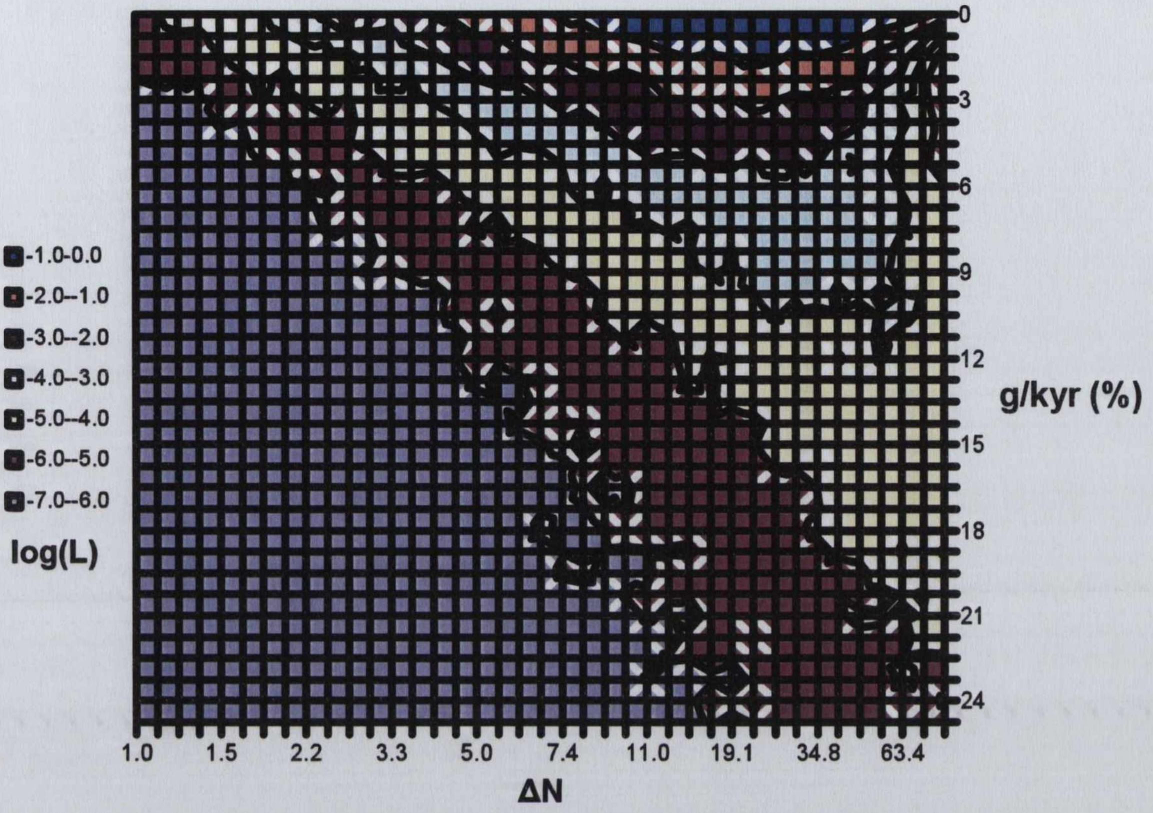
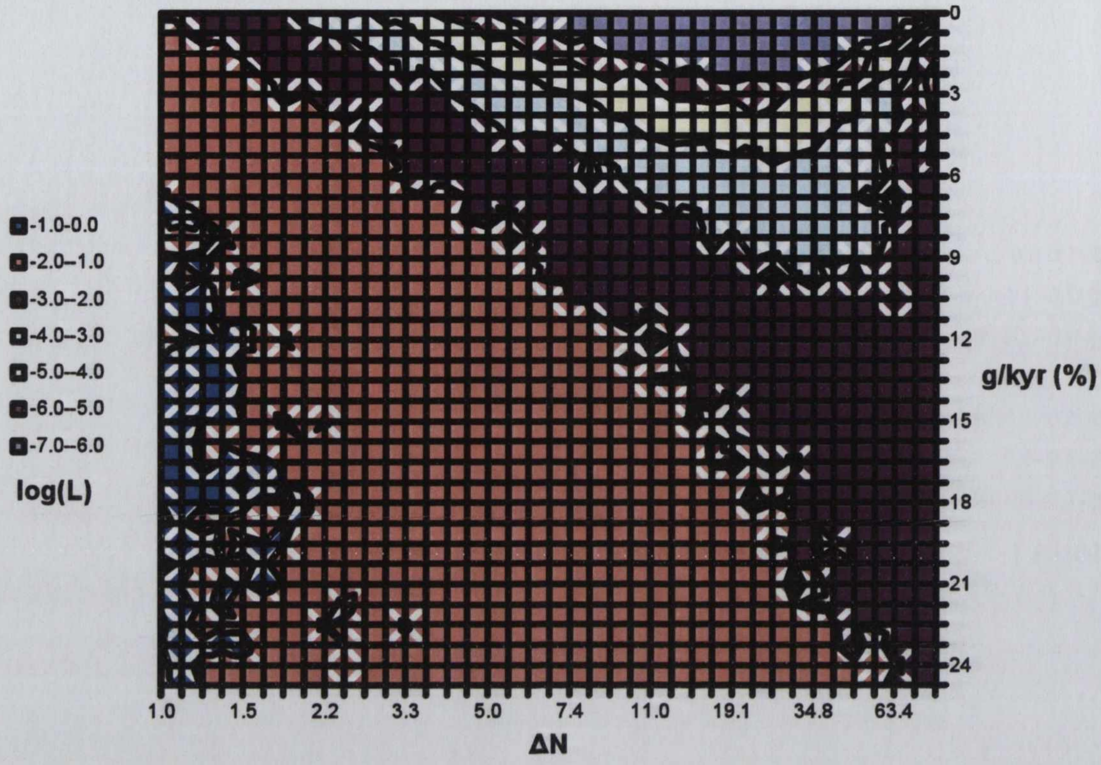


Table 7.8. Most likely simulation values for gene classes according to neutral, cytokine and TLR Tajima's *D* values.

Statistic	Genes	Observed	Average values for most likely parameters		
			Neutral	Cytokine	TLR
Tajima's *D*	Cytokines	0.572 (0.88)	-0.005 (0.02)	0.506 (0.19)	-0.312 (0.12)
	TLRs	-0.619 (0.95)	-0.007 (0.01)	0.503 (0.19)	-0.318 (0.12)
	All 18	-0.024 (1.02)	-0.006 (0.03)	0.504 (0.04)	-0.315 (0.03)
π[1]	Cytokines	10.26 (6.8)	7.65 (3.6)	8.71 (4.1)	6.97 (3.3)
	TLRs	11.62 (8.3)	13.75 (7.5)	15.70 (8.6)	12.56 (6.9)
	All 18	10.94 (7.2)	10.70 (6.7)	12.21 (7.7)	9.77 (6.2)
SNPs	Cytokines	30.22 (13.2)	34.2 (22.7)	34.0 (22.6)	34.2 (22.7)
	TLRs	54.67 (34.3)	51.2 (27.6)	51.0 (27.5)	51.4 (27.7)
	All 18	40.53 (27.4)	42.7 (29.0)	42.5 (28.8)	42.8 (29.0)

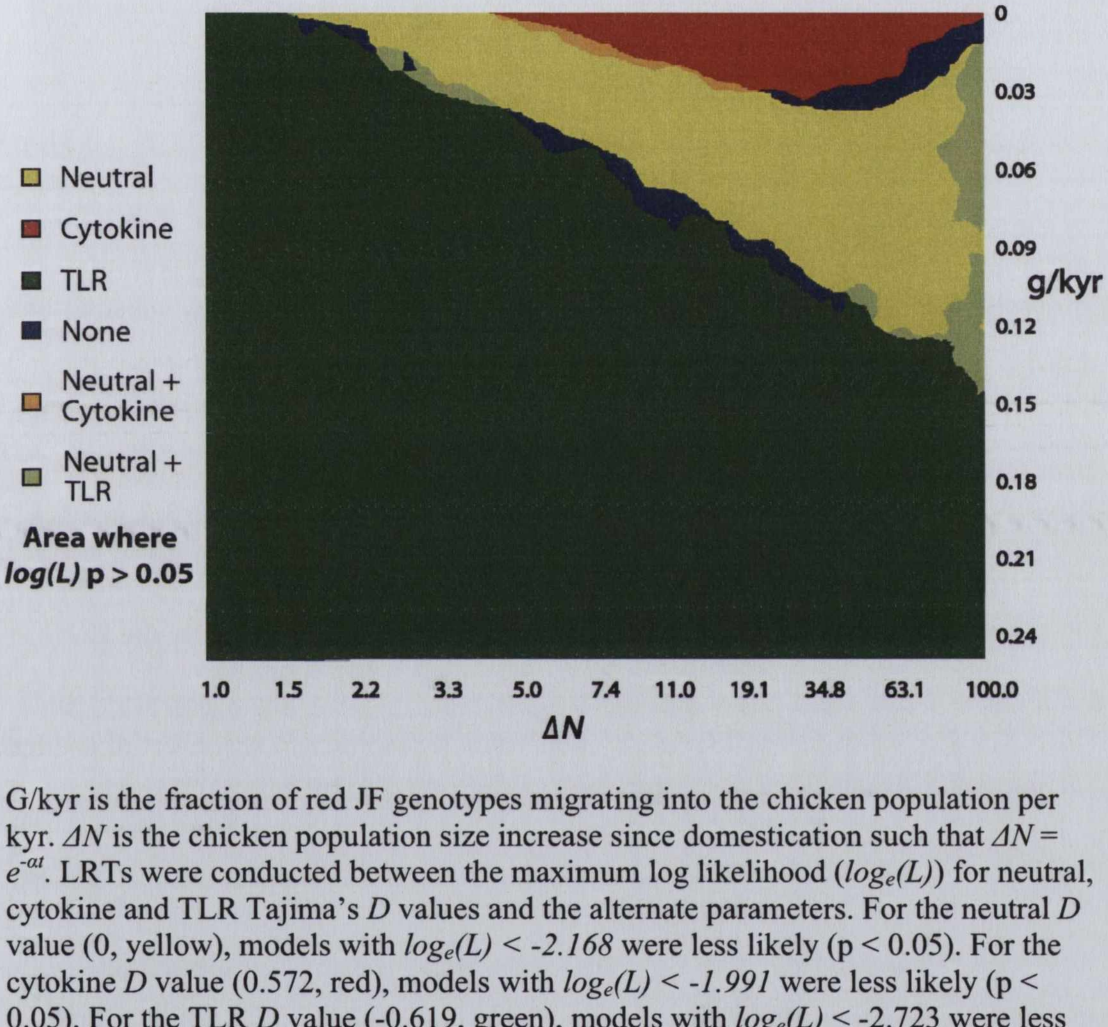
[1] π is measured per sequence rather than per kb. The standard deviations are in parentheses. The parameters value for the maximum likelihood models for the neutral Tajima's *D* value (0) had a 19.6-fold increase in population size (ΔN) and migration of red JF genotypes to the chicken population at a rate of 0.060 genotypes per kyr (g/kyr). The cytokine *D* (0.572) had a $\Delta N = 16.5$ and zero introgression. The TLR *D* (-0.619) had a constant population size and introgression of 0.165 g/kyr.

Figure 7.14. Contour map of the simulated TLR model (Tajima's $D = -0.619$) log likelihood ($log_e(L)$) for introgression of red JF genotypes into the chicken population (g/kyr %) and chicken population size increase (ΔN), models with $log_e(L) < -2.723$ were significantly less probable.



In addition, a number of initial simulations on the 18 resequenced genes were completed to test the neutrality of the demographic model design. These showed that with no ancestral population split, Tajima's *D* was neutral (-0.01). Interestingly, repeating this with $\Delta N = 10$ again produced a neutral *D* value (-0.01). When a chicken-red JF domestication split was introduced $0.1N_0$ generations ago, *D* became slightly negative (-0.16). Further tests showed that the timing of the domestication had no effect on *D*. However, for a population with a split, $\Delta N = 1$ and introgression = 0.05 g/kyr (meaning modern population was 60% chicken and 40% red JF), *D* became gradually more negative as the domestication became more ancient (-0.20 for event time $\leq 0.2N_0$; -0.40 for event time $\geq 0.4N_0$). This effect was even more pronounced for a population with a split, $\Delta N = 10$ and no introgression: *D* was negative for *0* to $0.1N_0$, positive for $0.2N_0$ to $0.6N_0$, and then gradually reverted to being highly negative by $1.2N_0$. Further tests on the neutrality of the demographic model that introduced additional ancestral red JF populations appeared to yield more negative Tajima's *D* values: the strength of this difference increased both with increasing number of populations and increasing age of the domestication event. These appeared to be synergistic such that more red populations increased the effect of the age, so that very ancient domestications (*$1.2N_0$*) from many red JF populations gave highly negative Tajima's *D* values (-1.30).

Figure 7.15. Significant LRT areas for the neutral, cytokine and TLR Tajima's *D* values.



G/kyr is the fraction of red JF genotypes migrating into the chicken population per kyr. *ΔN* is the chicken population size increase since domestication such that *$\Delta N = e^{-\alpha t}$*. LRTs were conducted between the maximum log likelihood (*$log_e(L)$*) for neutral, cytokine and TLR Tajima's *D* values and the alternate parameters. For the neutral *D* value (0, yellow), models with *$log_e(L) < -2.168$* were less likely ($p < 0.05$). For the cytokine *D* value (0.572, red), models with *$log_e(L) < -1.991$* were less likely ($p < 0.05$). For the TLR *D* value (-0.619, green), models with *$log_e(L)$* < -2.723 were less likely ($p < 0.05$). Areas where all LRTs had $p < 0.05$ are blue. Areas where parameters were not rejected with the neutral and cytokine *D* values are orange. Areas where parameters were consistent with the neutral and TLR *D* values are pale green.

7.3.5 GC content and gene conversion:

GC content for all genes was higher at coding (0.479; Table 7.9) than at noncoding sites (0.435). Gene-wide GC content was higher for cytokines (0.489) than for TLRs (0.414; Mann-Whitney *U* $p < 0.01$), and was also elevated at coding (0.516 vs 0.442 for TLR, $p < 0.01$) and at noncoding (0.484 vs 0.386 for TLR, $p < 0.01$) sites. In order to examine if the effect was an artefact of the genes' genomic positions, the gene GC values were compared to those of the chromosome on which they were located (Table 7.9; from Gao & Zhang 2006). The cytokine genes had GC content values 12% higher than the chromosomal average, whereas the TLRs were enriched by only 2%. At coding regions, the cytokines and TLRs had GC values enriched by a further 6% and 7%, respectively, showing that the GC elevation was genic for cytokines, and exonic for both gene groups. GC content was investigated further to determine if it affected the difference between the gene categories and observed diversity.

Increasing gene GC content correlated with a more positive Tajima's *D* ($r^2 = 0.31$, $p < 0.01$): this effect was stronger at silent sites (0.37, $p < 0.005$) but was weaker at CDS (0.06, ns). Protein-coding sequences are more likely to be subject to selective processes than silent sites because of the functional implications of any substitutions (Yang 2002); silent sites do not have this constraint and localised genomic phenomena may define their pattern of variation. When the correlation between GC content and Tajima's *D* effect was dissected between the two classes, it was far stronger at cytokines ($r^2 = 0.56$, $p = 0.01$) than it was at TLRs (0.12, ns). Again, this was largely a feature of silent sites ($r^2 = 0.41$, $p < 0.05$ for cytokines; 0.16, ns for TLRs) than of CDS (0.08, ns for cytokines; 0.07, ns for TLRs). Because the BGC tends to affect substitution across the gene rather than just at CDS (Webster et al. 2006), this effect may be a key factor in determining the variability observed at silent sites in the cytokine genes: the positive *D* may be caused by gene conversion.

An estimation of the number of gene conversion tracts between the broiler and heritage chickens showed that while the average rates for the groups were not different, the number of cytokine genes with tracts (8; Table 7.9) was higher than for the TLRs (4). Tajima's *D* correlated with the number of tracts for all genes ($r^2 = 0.18$, $p < 0.05$). When controlled for the number of bases affected by these events, the partial correlation was stronger (0.40, $p < 0.005$), and was significantly correlated for both cytokines (0.43, $p < 0.05$) and TLRs (0.46, $p < 0.025$). A correlation was evident between GC content and the number of gene conversion tracts for cytokines (0.36, $p < 0.05$) but not at TLRs (0.01, ns); this difference remained when the number of bases affected was controlled (0.35, $p < 0.05$ for cytokines; 0.12, ns for TLRs).

Table 7.9. Recombination, GC content, gene conversion and sites evolving neutrally for each gene.

Test/	Chr	R	R_M		GC content			NGC[1]	*f*[2]	B[3]
Gene				CDS	Noncoding	Gene	Chr			
GMCSF	13	45.5	4	0.488	0.506	0.503	0.443	1	0.235	2289
P value			*ns*							
IL12A	9	57.2	12	0.619	0.535	0.560	0.427	3	0.698	2466
P value			*0.010*							
IL13	13	43.6	19	0.589	0.546	0.552	0.443	2	0.086	313
P value			*<0.001*							
IL3	13	45.0	19	0.488	0.493	0.493	0.443	3	0	1445
P value			*<0.001*							
IL4	13	134.0	11	0.530	0.536	0.535	0.443	2	0.957	1320
P value			*0.022*							
IL5	13	140.0	9	0.481	0.449	0.450	0.443	1	0.683	745
P value			*ns*							
IL8	4	11.0	7	0.524	0.378	0.392	0.399	1	0	811
P value			*0.001*							
IL9	13	12.4	15	0.447	0.435	0.437	0.443	1	0.297	1104
P value			*<0.001*							
KK34	13	16.3	5	0.479	0.476	0.476	0.443	0	0.390	0
P value			*0.015*							
TLR15	3	111.0	9	0.437	0.372	0.427	0.398	0	0.298	0
P value			*0.032*							
TLR1LA	4	14.6	6	0.453	0.320	0.412	0.399	0	1.125	0
P value			*ns*							
TLR1LB	4	8.8	34	0.467	0.360	0.395	0.399	2	0.760	32
P value			*ns*							
TLR2B	4	10000	1	0.464	0.371	0.425	0.399	1	0.292	1056
P value			*ns*							
TLR2A	4	6.9	5	0.464	0.446	0.450	0.399	0	0.344	0
P value			*0.002*							
TLR3	4	42.3	41	0.375	0.388	0.383	0.399	5	0.353	6147
P value			*<0.001*							
TLR4	17	20.8	24	0.446	0.466	0.458	0.474	5	0.208	3508
P value			*<0.001*							
TLR5	3	0.001	4	0.393	0.362	0.387	0.398	0	0.358	0
P value			*<0.001*							
TLR7	1	12.3	13	0.397	0.377	0.386	0.398	0	0	0
P value			*<0.001*							

[1] Number of gene conversion tracts detected between broiler and heritage chickens. [2] Estimated fraction of neutral amino-acid changing substitutions. [3] Number of bases affected by gene conversion events. Statistics for which p > 0.05 are shown as "ns".

Adjacent genes on chr13, IL13 and IL4, both share significantly positive Tajima's *D* values and GC content elevated above the chromosomal level by 25% and 21%, respectively. A similar though less extreme pattern was observed for GMCSF and IL3, also on chr13, which had high *D* (1.30, 1.43) and local GC 14% and 11% above the chromosomal average, respectively. These results implicate gene conversion as a generator of allelic diversity at noncoding regions in cytokine genes.

7.3.6 Association of SNPs with disease:

A number of functionally relevant SNPs that had been previously associated with avian diseases, or had previously been identified in GeneView Report resequencing entries on GenBank were segregating in the chicken populations. The disease-associated SNPs were segregating in chicken at TLR4 (Table 7.10).

Table 7.10. List of nonsynonymous SNPs at TLR4 previously implicated in diseases that were segregating in the chicken populations.

Pos[1]	aA[2]	Major[3]	Minor[3]	Major[3]	Minor[3]	Major[3]	Minor[3]	Both[4]	f(CK)[5]	f(red)[5]
1863	83	Lys (K)	Glu (E)	AAA	GAA	-	H, G, V, C	B, R	0.07	0.13
3503	301	Asp (D)	Glu (E)	GAT	GAG	G, C, V	H	B, R	0.60	0.50
3628	343	Lys (K)	Arg (R)	AAA	AGA	G, C, V	H	B, R	0.63	0.88
3747	383	His (H)	Tyr (Y)	CAC	TAC	H, V	-	B, R, G, C	0.10	0.63
4432	611	Arg (R)	Gln (Q)	CGC	CAG	G, C, V	-	B, R, H	0.40	0.13

[1] Nucleotide position in the gene. [2] Amino acid position affected. [3] Major and minor alleles. [4] Groups with both alleles. [5] The SNPs' frequencies in the chicken and red JF populations are denoted as f(CK) and f(red). C stands for Ceylon JF, R for red JF, G for grey JF, V for green JF, B for broilers and H for heritage chickens. Site 83 was detailed by Ye et al. (2006) and Beaumont et al. (2003); sites 301, 343, 383 and 611 by Leveque et al. (2007).

K343R in TLR4 is associated with resistance and susceptibility to *SE* serovar Typhimurium (Leveque et al. 2003) and was segregating at a frequency of 0.63 among the chicken genotypes. It appeared to bisect a median-joining network of the gene (Figure 7.3-16) and the ancestral allele (K) was present in all the Ceylon, green and grey JF, and in only one red JF genotype. A K to R substitution is conservative: both are polar and positively charged – only the hydrophobicity is affected. This site is in the LRR of the extracellular domain and might be involved in ligand recognition (Leveque et al. 2003). Variation at K83 in TLR4 is implicated in resistance to *SE* serovar Enteriditis, as is a synonymous substitution at L73 (Beaumont et al. 2003). K83E was variable in the present study at low frequencies in the broiler (0.07) and red JF populations (0.13). The likely derived allele (K) was the same as the genome sequence and was variable only in the broiler and red JF. L73 at TLR4 (CTG to CTA, base 1835) was segregating in the chicken and red JF populations.

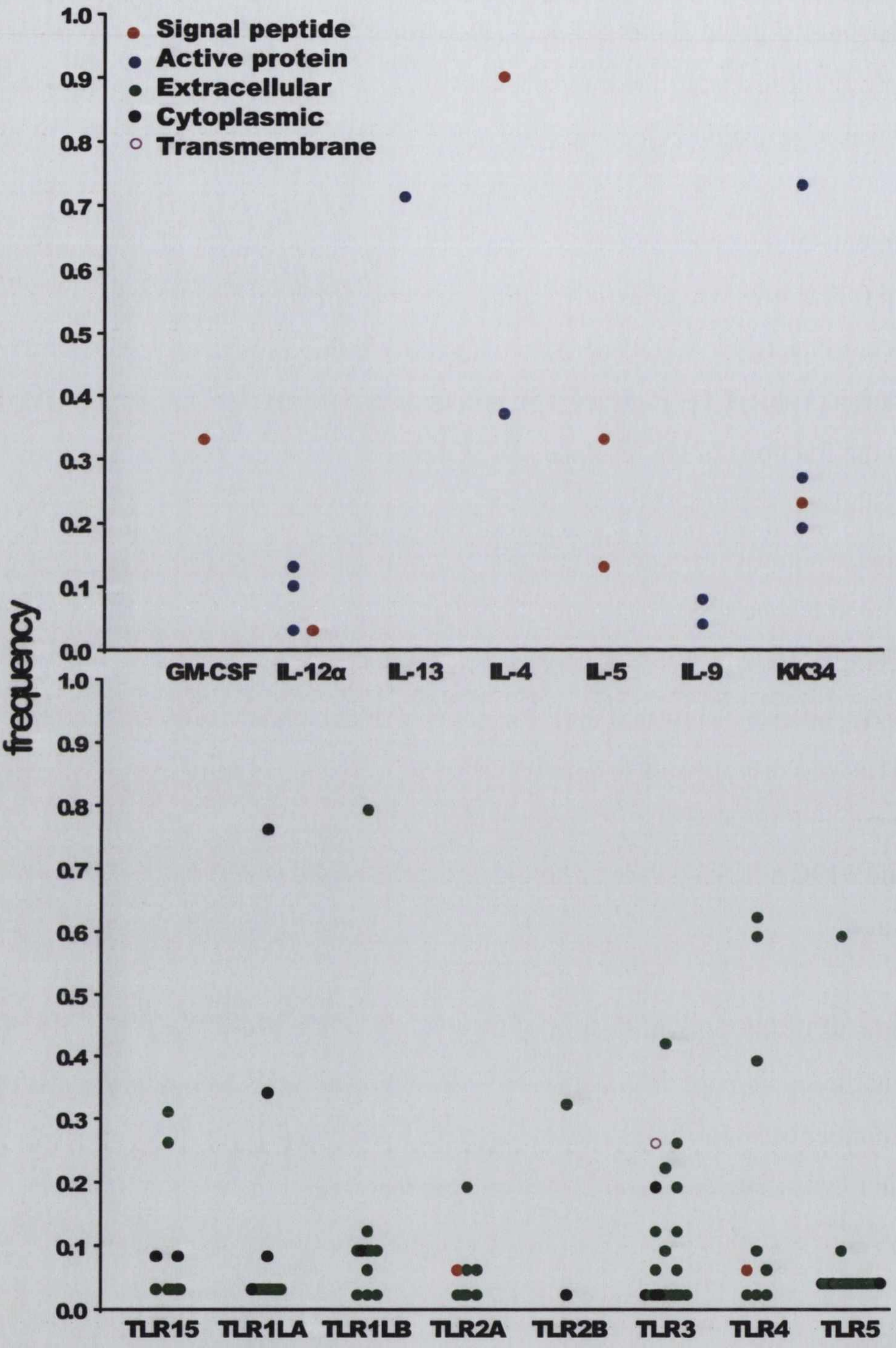
There were a large number of nonsynonymous SNPs segregating at intermediate frequencies or higher (> 0.1) in the chicken population that may be of functional relevance (Table 7.11). At TLR3, Q38K was segregating at a frequency of 0.23 in the chickens and 0.50 in red JF (Table 7.11). Q was present in all grey and Ceylon JF samples but the green JF possessed K, indicating that this site may be variable in multiple JF, obfuscating the true ancestral allele (Figure 7.3-15). Adjacent substitution N44S was segregating at identical frequencies in the chickens (0.23) and red JF (0.5).

Alignments of the chicken protein sequences with their human orthologs using T - Coffee (not shown) suggested that of the 29 replacement substitutions identified with an allele frequency > 0.1, 25 were in the extracellular domain, according to Uniprot annotation (Table 7.11, Figure 7.19), including those linked to disease at TLR4. Given the fractions of the resequenced TLR genes that were predicted to code for extracellular (58.7%; 3753 amino acids) and cytoplasmic (34.3%; 2194 amino acids) domains, a significant excess of protein changes occurred in the extracellular domain (one-tailed Fisher's Exact Test $p = 2.1 \times 10^{-5}$). The more variable DAF_N for the extracellular domain (0.126 ± 0.37 vs 0.142 ± 0.14 for the intracellular) indicated a higher abundance of low and high frequency alleles in that region. V788A at TLR1LA was orthologous to a human site in the TIR domain of the cytoplasmic polypeptide. R135K at IL12A was orthologous to a site in the active protein. M23T at IL4 and S19C at KK34 were orthologous to sites at the end of the signal peptide domain.

Alignments of the cytokines showed 7 (41%) nonsynonymous SNPs of 17 in total were in the signal peptide domain, which constituted 22% (193 amino acids) of the total number of amino acids resequenced; the remaining 10 (59%) were in the active protein (78%; 680 amino acids). There were marginally more replacement changes in the cytokine signal peptides than in the active domains ($p = 0.059$).

These patterns suggest that disease-associated polymorphic sites at TLRs may be those that can interact directly with PAMPs, though in the case of TLR1LA the substitution may be related to how this signal is transmitted to the TLR pathway through the TIR-recognising TRIF complex. For the cytokines, disease-associated SNPs appear to be more frequent in the signal peptide, a region that can dictate how often the protein precursor is activated in response to a signal.

Figure 7.19. Allele frequency of all nonsynonymous SNPs in the chicken samples according to functional domain.



The domains listed are: signal peptide (red), active protein (blue), extracellular (green), cytoplasmic (black) and transmembrane (unfilled pink).

Table 7.11. All nonsynonymous SNPs in the chicken samples.

Pos	aA	Ancestral	Derived	Ancestral	Derived	f(CK)	f(red)	Domain
GMCSF								
85	10	Ala (A)	Ser (S)	GCC	TCC	0.33	0	Signal peptide
IL12A								
90	4	His (H)	Asp (N)	CAC	AAC	0.03	0	Signal peptide
103	8	Ser (S)	Iso (I)	AGC	ATC	0.03	0	Signal peptide
1129	80	Val (V)	Iso (I)	GTC	ATC	0.03	0	Active protein
1478	135*	Arg (R)	Lys (K)	AGG	AAA	0.13	0	Active protein
1742	160	Pro (P)	Leu (L)	CCG	CTG	0.10	0	Active protein
IL13								
1260	46	Ala (A)	Val (V)	GCG	GTG	0.71	0.38	Active protein
IL4								
304	23*	Met (M)	Thr (T)	ATG	ACG	0.90	0.88	Signal peptide
522	49	Val (V)	Iso (I)	GTC	ATC	0.37	0	Active protein
IL5								
817	7	Leu (L)	Phe (F)	CTT	TTT	0.33	0	Signal peptide
832	12	Val (V)	Met (M)	GTG	ATG	0.13	0	Signal peptide
KK34								
772	19*	Ser (S)	Cys (C)	AGT	TGT	0.23	0.13	Signal peptide
805	32	Ser (S)	Arg (R)	AGC	CGC	0.27	0	Active protein
809	33	Met (M)	Thr (T)	ATG	ACG	0.73	1.00	Active protein
811/2	34	Asn (N)	Ala (A)	AAT	GCT	0.19	0	Active protein
IL9								
1493	102	Asn (N)	Tyr (Y)	AAC	TAC	0.04	0	Active protein
1536	116	Arg (R)	Gln (Q)	CGG	CAG	0.08	0	Active protein
TLR15								
652	148	Ala (A)	Thr (T)	GCT	ACT	0.04	0	Extracellular
815	202	Asn (N)	Iso (I)	AAT	ATT	0.04	0	Extracellular
920	237	Asn (N)	Iso (I)	AAC	ATC	0.04	0	Extracellular
1133	309	Ala (A)	Glu (E)	GCA	GAA	0.32	0.50	Extracellular
1171	321	Ser (S)	Cys (C)	AGT	TGT	0.04	0	Extracellular
1203	394	Leu (L)	Phe (F)	TTA	TTT	0.27	0.13	Extracellular
2294	695	Arg (R)	Lys (K)	AGG	AAG	0.09	0.25	Cytoplasmic
2809	867	Glu (E)	Lys (K)	GAA	AAA	0.09	0	Cytoplasmic
TLR1LA								
566	55	Leu (L)	Phe (F)	TTA	TTT	0.04	0	Extracellular
606	69	Thr (T)	Ser (S)	ACT	TCT	0.04	0	Extracellular
833	144	Leu (L)	Phe (F)	TTA	TTT	0.04	0	Extracellular
848	149	Leu (L)	Phe (F)	TTA	TTT	0.04	0.13	Extracellular
1011/2	204	Asn (N)	Phe (F)	AAT	TTT	0.04	0	Extracellular
1021	206	Leu (L)	Pro (P)	CTC	CCC	0.04	0	Extracellular
1025	208	Leu (L)	Phe (F)	TTA	TTT	0.04	0	Extracellular
1327	309	Ser (S)	Asn (N)	AGC	AAC	0.04	0	Extracellular
1372	324	Tyr (Y)	Phe (F)	TAT	TTT	0.04	0	Extracellular
1375	325	Tyr (Y)	Phe (F)	TAT	TTT	0.04	0	Extracellular
1477	359	Arg (R)	Pro (P)	CGA	CCA	0.04	0	Extracellular
1504	368	Ser (S)	Phe (F)	TCC	TTC	0.04	0	Extracellular
1586	395	Lys (K)	Asn (N)	AAA	AAT	0.04	0	Extracellular
2594	731	Leu (L)	Phe (F)	TTG	TTT	0.04	0	Cytoplasmic
2764	788*	Val (V)	Ala (A)	GTT	GCT	0.77	0.75	Cytoplasmic
Pos	aA	Ancestral	Derived	Ancestral	Derived	f(CK)	f(red)	Domain

2835	812	Iso (I)	Leu (L)	ATA	TTA	0.04	0	Cytoplasmic
2844	815	Cys (C)	Arg (R)	TGT	CGT	0.35	0.50	Cytoplasmic
2854	818	Lys (K)	Thr (T)	AAG	ACG	0.09	0.38	Cytoplasmic
TLR1LB								
3860	38	Pro (P)	Ser (S)	CCT	TCT	0.10	0.38	Extracellular
3882	45	Gly (G)	Asp (D)	GGT	GAT	0.13	0	Extracellular
3908	55	Pro (P)	Ser (S)	CCA	TCA	0.03	0.13	Extracellular
3920	59	Ala (A)	Thr (T)	GCA	ACA	0.07	0	Extracellular
3975	77	Thr (T)	Met (M)	ACG	ATG	0.03	0	Extracellular
4043	99	Phe (F)	Leu (L)	TTT	CTT	0.10	0	Extracellular
4065	106	Arg (R)	Glu (Q)	CGA	CAA	0.80	0.75	Extracellular
4103	119	Asp (D)	Asn (N)	GAT	AAT	0.10	0.25	Extracellular
4115	123	Val (V)	Iso (I)	GTA	ATA	0.10	0	Extracellular
5046	434	Leu (L)	Pro (P)	CTG	CCG	0.03	0	Extracellular
5654	637	Iso (I)	Val (V)	ATT	GTT	0.10	0.38	Cytoplasmic
TLR2B								
554	196	Val (V)	Leu (L)	GTG	CTG	0.33	0	Extracellular
2906	718	Glu (Q)	Lys (K)	CAA	AAA	0.03	0	Cytoplasmic
TLR2A								
5510	21	Tyr (Y)	Cys (C)	TAC	TGC	0.07	0	Signal Peptide
5645	66	Thr (T)	Met (M)	ACG	ATG	0.07	0	Extracellular
5701	85	Lys (K)	Glu (Q)	AAG	CAG	0.07	0.25	Extracellular
6642	398	Leu (L)	Phe (F)	TTA	TTT	0.03	0	Extracellular
6661	405	Lys (K)	Stop	AAA	TAA	0.03	0	Extracellular
6725	426	Asn (N)	Iso (I)	AAT	ATT	0.03	0	Extracellular
6996	516	Ser (S)	Arg (R)	AGC	AGA	0.20	0.50	Extracellular
6998	517	Arg (R)	Lys (K)	AGA	AAA	0.03	0	Extracellular
TLR3								
1776	11	Val (V)	Asn (N)	GTT	GAT	0.33	0.50	Extracellular
1847	38*	Lys (K)	Gln (Q)	AAA	CAA	0.23	0.50	Extracellular
1866	44*	Asn (N)	Ser (S)	AAT	AGT	0.23	0.50	Extracellular
3913	159	Thr (T)	Pro (P)	ACA	CCA	0.03	0	Extracellular
3992	186	Lys (K)	Thr (T)	AAA	ACA	0.13	0.13	Extracellular
4049	205	Lys (K)	Arg (R)	AAG	AGG	0.03	0	Extracellular
4273	280	Thr (T)	Ser (S)	ACT	TCT	0.43	0	Extracellular
4310	292	His (H)	Leu (L)	CAC	CTC	0.03	0	Extracellular
4322	296	Tyr (Y)	Phe (F)	TAC	TTC	0.03	0	Extracellular
4394	320	Tyr (Y)	Phe (F)	TAT	TTT	0.03	0	Extracellular
4470	345	Arg (R)	Ser (S)	AGG	AGC	0.27	0.63	Extracellular
4520	362	Gly (G)	Glu (E)	GGG	GAG	0.10	0.38	Extracellular
4811	459	Lys (K)	Arg (R)	AAG	AGG	0.13	0.13	Extracellular
5021	529	Asn (N)	Val (V)	GAT	GTT	0.03	0	Extracellular
5054	540	Ala (A)	Val (V)	GCG	GTG	0.13	0.13	Extracellular
5068	545	Asp (D)	His (H)	GAC	CAC	0.20	0.13	Extracellular
5209	592	Ala (A)	Ser (S)	GCT	TCT	0.20	0.13	Extracellular
5291/2	619	Lys (K)	Iso (I)	AAA	ATT	0.03	0	Extracellular
5381	649	Ala (A)	Val (V)	GCT	GTT	0.27	0.25	Transmembrane
5458	675	Iso (I)	Leu (K)	ATA	TTA	0.03	0	Cytoplasmic
5572	713	Thr (T)	Ser (S)	ACT	TCT	0.23	0.25	Cytoplasmic
5618	728	Glu (Q)	Gly (G)	GAA	GGA	0.03	0	Cytoplasmic
TLR4								
757	26	Ala (A)	Val (V)	GCA	GTA	0.07	0	Signal Peptide
Pos	aA	Ancestral	Derived	Ancestral	Derived	f(CK)	f(red)	Domain
1863	83*	Lys (K)	Glu (E)	AAA	GAA	0.07	0.13	Extracellular

3117	173	Lys (K)	Stop	AAG	TAG	0.03	0	Extracellular
3503	301*	Asp (D)	Glu (E)	GAT	GAG	0.60	0.50	Extracellular
3628	343*	Lys (K)	Arg (R)	AAA	AGA	0.63	0.88	Extracellular
3747	383*	His (H)	Tyr (Y)	CAC	TAC	0.10	0.63	Extracellular
3904	435	Tyr (Y)	Phe (F)	TAC	TTC	0.03	0	Extracellular
4021	474	Leu (L)	Pro (P)	CTC	CCC	0.03	0	Extracellular
4432	611*	Arg (R)	Gln (Q)	CGC	CAG	0.40	0.13	Extracellular
TLR5								
549	66	Ser (S)	Leu (L)	TCA	TTA	0.05	0	Extracellular
620	89	Lys (K)	Glu (E)	AAA	GAA	0.05	0	Extracellular
701	116	Phe (F)	Leu (L)	TTT	CTT	0.05	0	Extracellular
984	211	Tyr (Y)	Phe (F)	TAT	TTT	0.05	0	Extracellular
987	212	Arg (R)	Lys (K)	AGG	AAG	0.60	0.50	Extracellular
1562	403	Phe (F)	Leu (L)	TTC	CTC	0.05	0	Extracellular
1589	412	Ser (S)	Gly (G)	AGT	GGT	0.05	0	Extracellular
2010	553	Iso (I)	Thr (T)	ATA	ACA	0.09	1.00	Extracellular
2208	619	Ala (A)	Glu (E)	GCG	GAG	0.05	0	Extracellular
2585	744	Asn (N)	His (H)	AAT	CAT	0.05	0	Cytoplasmic
2885	844	Gln (Q)	Glu (E)	CAA	GAA	0.05	0	Cytoplasmic

Sites are listed below corresponding gene names. * Sites that are disease-associated or were in GeneView Report entries on GenBank. Pos stands for the nucleotide position in the gene. aA represents the amino acid position affected. The ancestral and derived alleles are denoted. C stands for Ceylon JF, R for red JF, G for grey JF, V for green JF, B for broilers and H for heritage chickens. The frequency of the SNPs in the chicken and red JF populations are denoted as f(CK) and f(red), respectively. The human amino acid domain orthologous to the chicken site is listed.

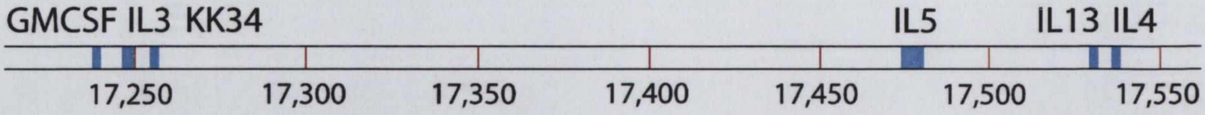
7.3.7 Variation within genes and gene clusters:

There are two cytokine clusters on chromosome 13: one at 17.23-17.26 Mb containing GMCSF, IL3 and KK34 and a second at 17.47-17.54 Mb for IL5, IL13 and IL4 (Figure 7.20, Table 7.12). These genes are likely to be ancient duplications of an ancestral gene (Avery et al. 2004). The first cluster showed a pattern of uniform high diversity. Tajima's *D* at GMCSF (1.30) and IL3 (1.43) were positive, indicative of balanced diversity, but that at KK34 was less so (0.37). IL3 had no mutations at nonsynonymous sites, but the levels of CDS polymorphism between GMCSF and KK34 within chickens ($\pi_A/\pi_S = 0.34$ for both) and between chickens and JF ($K_A/K_S = 0.40$ and 0.37, respectively) were similar.

At the second cytokine cluster, IL5 possessed a very high $[\pi_A/\pi_S]/[K_A/K_S]$ value (13.85, Table 7.12), likely to be a result of relaxed selective constraint; this contrasts strongly with the values at IL13 (0.65) and IL4 (0.76), which showed evidence of conservation within chicken. In fact, IL13 and IL4 had several shared characteristics in addition to $[\pi_A/\pi_S]/[K_A/K_S]$: significantly positive Tajima's *D* statistics (1.45 and

1.55, respectively), Fu and Li's D (1.72 and 1.74) and F (1.98 and 2.02) values, and significantly high numbers of recombination events (R_M).

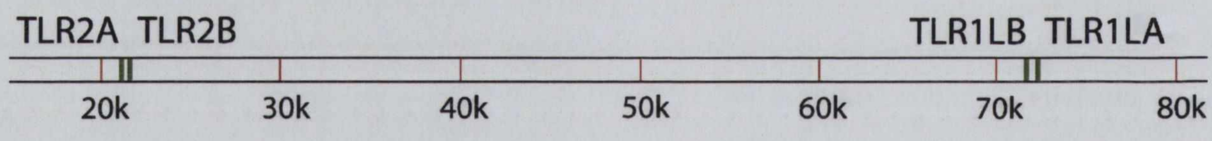
Figure 7.20. Gene map of two cytokine gene clusters chromosome 13.



Chromosome bases (red) are shown in units of 50 kb. Genes are shown in blue. Distances are approximately proportional. Gene sizes are not to scale.

There are two TLR gene clusters at chr4, both involving ancient duplications of TLR1 and TLR2 (Figure 7.21). TLR1LA had a significantly negative Tajima's D (-1.68; Table 7.12), a neutral Fay and Wu's H (-1.80) – these values contrast with TLR1LB's neutral D (-0.16), significant H (-23.13), though both genes did have similar π_A/π_S (0.54 and 0.63, respectively) and K_A/K_S values (0.40 and 0.48). TLR2A and B also showed divergent properties: TLR2A had a high Hd (0.94), a significantly negative Tajima's D (-1.41), a neutral rate of recombination ($R = 6.9$), and a low $[\pi_A/\pi_S]/[K_A/K_S]$ (0.44) in comparison to its π/K (1.05). By contrast, TLR2B had very low diversity ($Hd = 0.77$), a positive Tajima's D (0.60), very high recombination and $[\pi_A/\pi_S]/[K_A/K_S]$ (1.36) greater than its gene-wide π/K (0.32).

Figure 7.21. Gene map of two TLR gene clusters chromosome 4.



Chromosome bases (red) are shown in units of 10 Mb. Genes are shown in green. Distances are approximately proportional. Gene sizes are not to scale.

IL8 and TLR3 are also located on chr4 and IL9 is on chr13. Variation at IL8 was unusual for a cytokine: it had no cSNPs, and though its network was indicative of balanced diversity, its Tajima's D was negative (-0.46). Similarly, TLR3's D value was negative (-0.42); however this gene had a large number of intermediate- and high-frequency nonsynonymous SNP alleles, two of which were segregating in complete LD (K38Q and N44S). TLR3 also had a higher number than average of detectable gene conversion events (5), implying a possible role for this effect in modulating diversity. IL9 had extensive allelic variability (*$Hd = 0.99$*), likely to be driven by significantly high recombination ($R_M = 15$). At nonsynonymous sites IL9 had low divergence between chicken and JF but high variation within the chicken population ($[\pi_A/\pi_S]/[K_A/K_S] = 2.43$).

TLR5 and TLR15 are both located on chr3. TLR5 had a significantly negative Tajima's *D* (-2.04), consistent with directional selection and had a high-frequency segregating nonsynonymous SNP (R212K) that bisected its haplotype network (Figure 7.3-17). TLR15 had an amino acid-changing mutation (A309E) at an intermediate frequency and had high levels of haplotype variation ($Hd = 0.99$).

On chr17, TLR4 had a number of nonsynonymous SNPs that bisected the network: D301E, K343R and R611Q. This gene also showed high allelic diversity ($Hd = 0.95$), a high number of recombinants ($R_M = 24$), and a positive Tajima's *D* (0.42). On chr1, TLR7 also had significantly high recombination ($R_M = 13$) and elevated diversity (*Hd* = 0.99), however this gene had no nonsynonymous mutations in the chicken population.

7.3.8 Diversity at MC1R:

F_{ST} values indicated significant differentiation between chicken and red (0.885, p = 0.045), grey (0.780, p = 0.036) and Ceylon JF (0.874, p < 0.001) but not between broilers and heritage (0.078). F_{ST} values among the JF tended to be lower (red-grey, 0.167; red-Ceylon, 0.875; grey-Ceylon, 0.665), suggesting a genetic division of the chicken and JF groups. Using Arlequin version 2.001 (Schneider et al. 2000), AMOVA assigned components of total variation to different levels of population and group structure (Excoffier et al. 1992). No variation partitioned between broilers and heritage chickens, whereas 0.29 did between red, grey and Ceylon JF, and 0.56 between the chickens and JF, indicating that variability between the JF species was comparable to that between the JF and chicken.

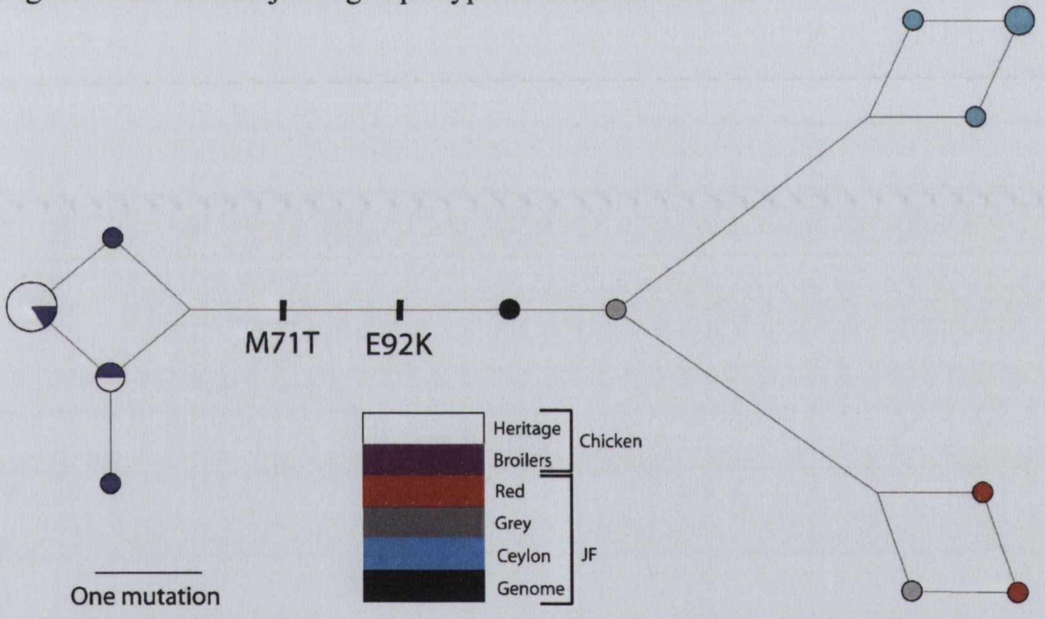
Table 7.12. Key gene characteristics from Tables 7.6, 7.7 and 7.8 listed according to their genomic positions for genes not separated by more than 50 kb. Approximate gene start ("Start") and end ("End") points are listed according to chromosomal positions in kilobases. Genes in pale green are those in cytokine cluster 1; those in orange cytokine cluster 2; those in purple the TLR1 cluster; and those in pale blue the TLR2 cluster.

Gene	Chr	Start	End	π per kb	θ_W per kb	Tajima's D	Fu & Li's D	Fu & Li's F	R
GMCSF	13	17,234	17,237	1.93	1.39	1.299	0.794	1.158	45.5
P value				*ns*	*ns*	*0.017*	*ns*	*ns*	
IL3		17,245	17,250	2.67	1.93	1.427	1.267	1.616	45.0
P value				*ns*	*ns*	*<0.001*	*0.015*	*ns*	
KK34		17,255	17,257	1.13	1.02	0.370	1.349	1.238	16.3
P value				*ns*	*ns*	*ns*	*0.031*	*ns*	
IL5	13	17,472	17,483	2.42	2.71	-0.389	0.136	-0.061	140.0
P value				*ns*	*ns*	*ns*	*ns*	*ns*	
IL13		17,529	17,532	3.98	2.85	1.454	1.720	1.979	43.6
P value				*ns*	*0.054*	*<0.001*	*<0.001*	*<0.001*	
IL4		17,535	15,537	3.12	2.15	1.551	1.738	2.020	134.0
P value				*ns*	*ns*	*0.001*	*0.001*	*<0.001*	
TLR1LB	4	71,549	71,555	3.20	3.34	-0.159	1.444	1.036	8.8
P value				*ns*	*ns*	*ns*	*0.010*	*ns*	
TLR1LA		71,561	71,567	1.68	3.02	-1.679	-3.210	-3.251	14.6
P value				*ns*	*0.010*	*0.002*	*<0.001*	*<0.001*	
TLR2A	4	21,101	21,108	0.70	1.13	-1.405	-0.859	-1.287	6.9
P value				*ns*	*0.031*	*0.003*	*ns*	*ns*	
TLR2B		21,109	21,116	0.33	0.27	0.601	0.010	0.217	>10000
P value				*ns*	*ns*	*ns*	*ns*	*ns*	

Gene	Chr	Start	End	R_M	π_A/π_S	K_A/K_S	$\frac{\pi_A/\pi_S}{K_A/K_S}$	K per kb	π/K
GMCSF	13	17,234	17,237	4	0.338	0.341	0.991	4.420	0.437
P value				*ns*					
IL3		17,245	17,250	19	0	0	0	3.550	0.752
P value				*<0.001*					
KK34		17,255	17,257	5	0.396	0.367	1.079	1.990	0.568
P value				*0.015*					
IL5	13	17,472	17,483	9	3.462	0.250	13.848	4.610	0.525
P value				*ns*					
IL13		17,529	17,532	19	0.108	0.166	0.651	5.990	0.664
P value				*<0.001*					
IL4		17,535	15,537	11	0.636	0.835	0.762	9.130	0.342
P value				*0.022*					
TLR1LB	4	71,549	71,555	34	0.629	0.478	1.316	3.620	0.884
P value				*ns*					
TLR1LA		71,561	71,567	6	0.540	0.403	1.340	2.300	0.730
P value				*ns*					
TLR2A	4	21,101	21,108	5	0.230	0.529	0.435	0.670	1.045
P value				*0.002*					
TLR2B		21,109	21,116	1	0.347	0.256	1.355	1.040	0.317
P value				*ns*					

A total of 15 SNPs in chicken and JF samples were identified at MC1R, six of which were coding – two of these were nonsynonymous. Both of these replacement mutations (M71T at base T285C and E92K at base G347A) separated the chicken and JF (Figure 7.22) in a median-joining network. An additional synonymous SNP (N23) at base 142 also separated the JF and chicken. The absence of nonsynonymous SNPs within each of the chicken or JF groups suggested that the coding portion of the gene was conserved, implying functional relevance for the substitutions at sites 71 and 92. Sites 23 and 92 are in extracellular domains and site 71 is in a transmembrane domain (Uniprot entry P55167). Predicted impacts of the substitutions by PMut (Ferrer-Costa et al. 2005) and SIFT (Ng and Henikoff 2003) indicated that E92K and M71T were unlikely to affect protein structure (Table 7.13).

Figure 7.22. Median-joining haplotype network for MC1R.



Broilers are purple, heritage chickens are white, red JF are red, Ceylon JF are blue, grey JF are grey and the genome sequence is black. Branch lengths are proportional to the mutational distance listed. Nonsynonymous SNPs are shown.

The chicken MC1R samples had four haplotypes separated by three SNPs located in the 5' and 3' UTR. Variation was low, highlighting further the sequence conservation (*$Hd = 0.644$*, *$\pi = 2.21$* per kb, *$\theta_W = 2.69$* per kb); other descriptive statistics were neutral (Tajima's $D = -0.66$, Fu's $F_S = -1.18$, Fay & Wu's $H = -0.27$, Fu & Li's $D = -1.02$, Fu & Li's $F = -1.10$, $R_M = 1$) but divergence between chicken and JF was comparatively high ($K = 21.0$ per kb), particularly in comparison to the cytokine and

TLR genes. GC content (0.620) was 50% higher than the chromosomal value for chr11 (0.414; Gao & Zhang 2006), this was a gene-wide effect: GC at coding sites (0.646) was about the same at noncoding sites (0.609).

Table 7.13. Protein impact predictions for nonsynonymous substitutions.

Software	Substitution	M71T	E92K
PMut	Prediction	neutral	neutral
	Score	0.310	0.183
	Reliability value	3	6
SIFT	Prediction	tolerated	tolerated
	Score	1.00	1.00

The PMut score indicates more acceptability in substitution type as it decreases. The reliability value increases with higher confidence in the prediction (from 0 to 9). The SIFT score predicts the substitutions would not be tolerated if the score < 0.05. The median sequence information value, at 3.17, was less than 3.25, indicating the predictions were probably reliable.

Using the PAML 3.15 package (Yang 1997), a LRT performed between the branch models log likelihoods (L) shows the variable model (L = -416.8) was significantly more probable than the neutral one (L = -440.9, $p < 1 \times 10^{-4}$). ω estimated for the chicken branch (0.142) was significantly higher than that for other JF (ω = 0.0001), consistent with a faster evolutionary rate in chicken compared to the JF.

7.4 Discussion

7.4.1 The origin of high diversity in chicken:

An emergent pattern of high allelic diversity at the TLR and cytokine genes has been observed at other chicken DNA regions (Liu et al. 2006, Worley et al. 2008, Berlin et al. 2008, Muchadeyi et al. 2008, Berthouly et al. 2009). The chicken has had a complex demographic history since an initial series of domestication events from geographically separated red JF in South and South-East Asia (West & Zhou 1989, Fumihito et al. 1994, Fumihito et al. 1996, Liu et al. 2006, Oka et al. 2007, Razafindraibe et al. 2008). This was complicated by further inferred introgressions of red and other JF, preventing genetic separation of chicken from red JF and further enhancing chicken variability (Eriksson et al. 2008, Silva et al. 2009, Nishibori et al. 2005). These domestication events would have also resulted in changes in the chicken's population structure and density, and the widespread migration and trading of genetically diverse chickens would have admixed divergent types (for example, Storey et al. 2007). This combination of multiple separate genetic origins, human-driven admixture and ease of portability appears to have caused the trend of high diversity and minimal geographic structuring of modern chicken populations.

This study presents clear evidence that red JF and chickens have not been genetically isolated since domestication, as suggested elsewhere (International Chicken Genome Sequencing Consortium 2004, Kanginakudru et al. 2008, Bao et al. 2008, Granevitze et al. 2009). Previous studies have found evidence for introgression of grey JF, which would have further altered the genetic variation present in the domestic chicken. Here, grey, Ceylon and green JF separated from chicken and red JF into phylogenetically distinct groups, suggesting at least some separation of genetic histories. However, at two genes (TLR1LA and TLR2A) there was no clear division of variability between the chicken, red and grey JF populations: these likely represent new examples of introgression of grey JF into the chicken population in addition to that demonstrated for the yellow skin locus (Eriksson et al. 2008).

7.4.2 Shared adaptive patterns at gene clusters:

Certain chicken MHC genes that are co-expressed in response to infection could be co-regulated (Ortutay & Vihinen 2006), so shared patterns of variation and evolutionary history identified here may serve as indicators of common functional roles at clustered genes.

All of the cytokine genes grouped on chr13 are likely to be duplications of an ancestral gene (Avery et al. 2004) – these genes separate into two tight groups. At the first cluster, IL3 was likely to be under strong purifying selection at coding sites, though like GMCSF it did possess a signature of balanced diversity – both may be subject to BGC. The other gene in this cluster, KK34, is a cytokine-like transcript that is a homologue of IL5 (Koskela et al. 2004): the function of IL5's gene product may have changed so KK34 perform part of its former roles (Avery et al. 2004). This is likely to have altered variation at KK34 and IL5 so that the genetic history of each is unique compared to the norm for the chromosomal region. The differing evolutionary trends at IL5 compared to the other gene pair in the cluster support the hypothesis that the function of IL5 may have changed in the avian lineage.

At the second cytokine gene cluster that is involved in initialising the T_H2 response (Kaneko et al. 2007), IL13 and IL4 shared a common pattern of variation, so it seems possible that both have undergone similar selective processes, including possible BGC. Interleukins 4 and 13 may have shared functional roles: for example, they both respond to Marek's disease virus (Heidari et al. 2008), and in humans both mediate B cell activity and immunoglobulin production (Coffman et al. 1986, McKenzie et al. 1993), so it appears possible that IL13 and IL4 are positioned in an adjacent manner in order to optimise co-expression. This is present in mice (Guo et al. 2005), but IL4 expression varies in tandem with chromatin accessibility (Agarwal & Rao 1998, Guo et al. 2002).

The other adjacent locus in the cluster, IL5, had very differing values for Tajima's *D* as well as Fu and Li's *D* and *F* compared to those for IL4 and IL13. This difference can be explained in part by their differing transcription directions (Nakayama et al. 2005): intergenic regions between IL4 and IL13 may be transcribed (Baguet et al. 2005); and also by a separate local histone hyperacetylation region for chromatin remodelling at IL5 to that upstream of IL13 for IL13 and IL4 in humans (Yamashita et al. 2002). These discrete nucleosomes are key determinants of human IL13, IL4 and IL5 expression kinetics (Yamashita et al. 2004) and act to differentiate the relative levels of IL4 produced by CD8 and CD4+ T cells (Nakayama et al. 2005). Taken with the shared evolutionary signatures between IL13 and IL4, this highlights their likely co-expression in chicken, and suggests that this is related to pathogen-driven selective pressure on the T_H2 component of the adaptive immune response.

The TLR1 and TLR2 gene products on chr4 respond to *Mycobacterium avium* infection (Higuchi et al. 2008), so one reason for their close proximity may be co-expression. At the TLR1 cluster, the duplication of TLR1LA and TLR1LB 147 mya (Temperly et al. 2008) may have caused functionalisation of at least one of these genes because summary statistics indicated clear differences between them. The likelihood of non-functionalisation of a duplicate gene copy decreases with increasing N_e, so for the chicken, the probability of neo-functionalisation, which may have occurred for the TLR1 and TLR2 gene pairs (Conery and Lynch 2000). The extracellular portion of TLR1LA has previously shown evidence of directional selection (Yilmaz et al. 2005): this was supported by the significantly negative Tajima's *D* here. TLR1LB too has been suggested as a candidate for directional selection (Yilmaz et al. 2005), and it covers the functions of both mammalian TLR1 and TLR6 and can interact with TLR2A through its central LRR region (Keestra et al. 2007). At TLR1LA, it is possible that recent introgression of a grey JF gene variant has been positively selected, giving rise to the negative Tajima's *D*, which was also seen at the other gene with evidence of grey JF introgression, TLR2A. TLR2A and B are the result of a duplication 65 mya (Temperly et al. 2008). Summary statistics suggest both genes were subject to purifying selection (Asthana et al. 2007). However, TLR2B had evidence of balanced diversity and had one intermediate-frequency nonsynonymous SNP (V196L). On the other hand, TLR2A's high silent site diversity and negative Tajima's *D* supports a hypothesis of directional selection, which is supported by evidence elsewhere (Cormican et al. 2009). This selection signal may have been modulated by extensive haplotype variability due to possible introgression of grey JF, and also by an intermediate frequency nonsynonymous SNP (S516R) that bisected the network.

7.4.3 Specific gene histories:

Like IL5, a very high $[\pi_A/\pi_S]/[K_A/K_S]$ value (18.61) was observed at the IL12A locus on chr9, however it has a high degree of functional similarity with that of higher vertebrates and is likely to be conserved since the ancient split of birds and mammals (Degen et al. 2004): hence IL12A may be undergoing rapid adaptation in chicken rather than pseudogenisation like IL5, which also had a very high $[\pi_A/\pi_S]/[K_A/K_S]$ value. Further support from this comes from two nonsynonymous SNPs (R135K and P160L) that bisected the IL12A haplotype network.

TLR3 had a large number of intermediate- and high-frequency nonsynonymous SNP alleles, two of which were adjacent (K38Q and N44S); diversity at this gene may also be subject to BGC. In tests with mice, TLR3 was associated with protection against H5N1 avian influenza (Wong et al. 2009). TLR5 showed properties consistent with directional selection and had one high-frequency nonsynonymous SNP (212K) – this gene is under directional selection in primates (Wlasiuk et al. 2009). An ancient duplicate of the ancestral TLR1/TLR2 gene (Lynn et al. 2003), TLR15, also had a functional substitution mutation (A309E) that appeared to divide variability at this gene. Initial identification of TLRs 3, 5 and 7 suggested these genes are undergoing purifying selection (Yilmaz et al. 2005). TLR4 had a number of nonsynonymous SNPs that were segregating at moderate or high frequencies and have been associated with chicken diseases (D301E, K343R and R611Q). These mutations may represent an artefact of the selective processes endured by different ancient chicken populations that have been retained to optimise immunity to a wide range of pathogens.

7.4.4 Adaptive evolution at the MC1R gene:

This chapter identified functional substitutions at MC1R that phylogenetically separate chickens from red, grey and Ceylon JF, despite a trend of low diversity at the gene. This contrasts strongly with the other chicken and JF genes, where diversity was shared between red JF and chicken. This implies a role for human preference favouring birds with lighter feathers during domestication of the chicken from its wild JF ancestors, which would have been shaped by the ancient evolution of pigmentation in avian ancestors dating from at least the Middle Eocene, about 40-50 mya (Vinther et al. 2009). The mutations detected here that divided the chickens from the JF (N23, M71T and E92K) were previously observed to be associated with plumage and skin colour in Chinese village chickens (Yang et al. 2008). The K92 variant present in red JF constitutively activates MC1R without need for α-MSH (Takeuchi et al. 1996b); E92 is characteristic of domestic chickens (Kerje et al. 2003). The colouration pattern matches that of the bananaquit (*Coereba flaveola;* Theron et al. 2001), lesser snow goose (*Chen caerulescens*), arctic skua (*Stercorarius parasiticus*; Mundy 2005) and quail (Minvielle et al. 2009). Site 92 is evolving in other avian species: it segregates as E92A in the black-necked swan (*Cygnus melanocoryphus*; Pointer & Mundy 2008). E92 is fixed in other birds, like the blue-crowned manakin (*Lepidothrix coronata*; Cheviron et al. 2006). The evolutionary history of site 71 is less clear: M but not T is present in quail (*Coturnix japonica*; Nadeau et al. 2006).

Alignments of chicken and mammalian MC1R protein sequences showed that E92 and M71 are conserved (Jackson 1997). A number of substitutions at MC1R are shared between mammals, birds and reptiles (Rosenblum et al. 2004): chicken shares K92 with the mouse, in which it is orthologous to site 94, indicating the possibility that the E to K change was present in the avian-mammal ancestor (Ling et al. 2003). Other mutations at MC1R generate different phenotypic variation in many species: the swan (Pointer & Mundy 2008), red-footed booby (*Sula sula*; Baião et al. 2007), flycatcher (*Monarcha castaneiventris*; Uy et al. 2009), human (Valverde et al. 1995, Bamshad & Wooding 2003), cow (*Bos*; Mohanty et al. 2008), rabbit (*Oryctolagus cuniculus*; Fontanesi et al. 2006), sheep (*Ovis aries*; Deng et al. 2009), and goat (*Capra aegagrus*; Wu et al. 2006). The latter possess an E226K change associated with coat colour in the cytoplasmic domain (Uniprot entry P56444). Additionally, a 3' MC1R splice variant has been observed in humans but not yet in other vertebrates (Tan et al. 1999).

This analysis follows a trend of intentional human selective pressure at MC1R for colouration types during domestication, like that in the domestic pig (*Sus scrofa*; Fang et al. 2009). This effect could be combined with naturally occurring sexual selection for plumage colour: polymorphisms at MC1R display sexual dimorphism and incomplete dominance (Kerje et al. 2003). Adaptation at the gene is probably shaped by immunological roles: a bacterium commonly found on bird feathers (*Bacillus licheniformis*) degrades lighter feathers more quickly than darker ones (Goldstein et al. 2004). β-defensins signal through MC1R to induce coat colour variation in dogs (*Canis lupus*; Candille et al. 2007, Anderson et al. 2009). Plumage intensity is associated with the power of the immune response in tawny owls (*Strix aluco*; Gasparini et al. 2009), barn owls (*Tyto alba*; Roulin et al. 2000, 2001) and greenfinches (*Carduelis chloris*; Saks et al. 2003). MC1R has a role in mediating the anti-inflammatory effect of α-MSH (Schiöth et al. 2003); other melanocortin receptors function in metabolism (Andersson 2003) and the nervous system (Takeuchi et al. 1999). In wild populations, an enhanced immune response may have had an associated effect of elevating metabolism, which would have led to selective breeding for domestic birds whose energy was invested in traits beneficial to humans. Behavioural studies indicate that while lighter-coloured domestic animals may be more docile and compliant (Ducrest et al. 2008): redder widowbirds are typically more aggressive (*Euplectes ardens*; Pryke et al. 2002) and red Gouldian finches may be innately intimidating to other birds (*Erythrura gouldiae*; Pryke 2009), which would have a negative impact on flock output in chickens. Consequently, plumage-associated variation at MC1R may have historically served as a marker for traits linked to successful farming, giving rise to the genetic signature apparent in modern birds. Nonetheless, it cannot yet be discounted that humans may have selectively bred chickens based on plumage colour for purely aesthetic reasons.

7.4.5 Evolutionary differences between gene classes:

Analysis of nucleotide diversity and Tajima's *D* indicated that chicken cytokine and TLR gene classes have different evolutionary signatures. As a consequence of having larger π values, the cytokines had more positive Tajima's *D* on average, suggestive of balancing selection (Tajima 1983b). In contrast, the TLR family tended towards negative *D* values, which is more consistent with positive or negative selection. The genomic rate of segregating mutations likely to be disadvantageous (0.20; Axellson & Ellegren 2009) suggested this fraction was relatively neutral at cytokines (0.21) but was high at TLRs (0.37; Liti et al. 2009), signifying that the negative Tajima's *D* trend at TLRs was not caused by purifying selection.

The receptor and mediator allele frequency spectra indicated that at intermediate DAFs there was a substantial excess of nonsynonymous substitutions in cytokines compared with TLRs, even though the rates at synonymous sites were about the same for both. This surfeit of variation at nonsynonymous sites contributed to the more positive Tajima's *D* value present in the cytokine group because the difference was present at intermediate but not at low DAF. This suggests that the balanced pattern of variation at cytokines has functional relevance.

On the basis that the contrasting modes of evolution at these gene classes were likely to have been shaped by the demographic effects of domestication, coalescent simulations varied rates of chicken population size expansion and introgression of red JF into the chicken population. Domestication was simulated as an event dividing an ancient *Gallus gallus* population into two groups: an expanding chicken population and a red JF population, some of which migrated into the chicken population. The maximum likelihoods for three Tajima's *D* values (neutral, $D = 0$; cytokine, $D = 0.57$; and TLR, $D = -0.62$) were determined and LRTs were conducted between these and the alternative models. The three scenarios were simulated for a range of models that varied the rate of chicken population size expansion and introgression of red JF per generation. These effects could not explain both the cytokine and TLR *D* values ($p = 0.014$) and could only marginally explain the differences between the neutral scenario and cytokine ($p = 0.062$) and neutral and TLR ($p = 0.057$) values. Even when the scenarios were designed to explain as much as possible of the variation at each gene class, the simulated π per gene for the cytokines (8.71) was much lower than the observed rate (10.26), and the simulated value for the TLRs (12.56) was still higher than the observed one (11.62). This shows that the demographic history of the chicken does not appear to explain sufficiently the differences observed in the resequenced gene classes.

Further differences emerged between the gene classes in relation to this and possible BGC. GC content was elevated at cytokine genes and this correlated positively with Tajima's *D* values for cytokines, but less well for TLRs. This link between GC content and *D* was stronger when only silent sites were included: although coding regions showed higher GC, little correlation with *D* was observed. Gene conversion correlated with *D* and GC content for the cytokine class but only with *D* for the TLRs. Although GC content may mimic directional selection at coding sequences (Hurst 2009) and thus result in bias (Berglund et al. 2009), the correlation with Tajima's *D* suggests that novel alleles created by BGC could increase to intermediate frequencies at cytokine genes. Locally high levels of GC content, which may be characteristic of chicken immune genes, were present at cytokines here and may contribute to net higher diversity (Spencer *et al.* 2006).

Much variation at chicken immune genes that was originally generated by multiple origins and admixture (Liu et al. 2006, Bao et al. 2008) appears to have caused a trend of high diversity at the genes resequenced in this study. However, it does not sufficiently account for the distinctive patterns of variation observed between the cytokines and TLRs, and implies selective processes at one or both of these functional categories.

7.4.6 The causes of the differences between gene classes:

Among the multitude of effects of domestication were higher population densities as well as challenges by novel pathogens in new and variable environments, and from other domesticated animals in addition to other chicken and red JF flocks. These effects are likely to have initiated new adaptive forces on host defence genes as the immune system evolved to counter novel diseases. Much of the diversity generated by the chicken's population history continues to be maintained at disease-associated genes (Worley et al. 2008, Berlin et al. 2008, Downing et al. 2009a, Downing et al. 2009b, Downing et al. 2009c), indicating that the selective processes acting on immune defence genes result in the preservation of this variability.

One possibility explaining the persistence of elevated diversity is that standing variation provides a genetic resource for combating novel challenges and thus a form of hybrid vigour in the face of pathogen attacks. Selection acting on previously neutral alleles segregating at non-singleton frequencies in the population will result in signatures different to those from advantageous *de novo* mutations (Innan & Kim 2004, Hermisson & Pennings 2005, Przeworski et al. 2005). Interestingly, immunity-related multilocus heterozygous advantage has been demonstrated in humans (Lyons et al. 2009a, 2009b) and may be present in avian species as well (Reid et al. 2007, Mulard et al. 2009). However, these factors do not explain the fundamental evolutionary differences between the mediator and receptor gene classes.

TLR genes encode transmembrane receptors whose functions are to recognise PAMPs through the extracellular domain, and initiate innate and adaptive immune mechanisms appropriate to the response required with the cytoplasmic domain, including activating pro-inflammatory cytokine expression (Zhou et al. 2007).

Adaptive pressures are likely to be more distinct at TLRs than cytokines because, as sentinels of the immune system, TLRs are the first immune proteins that can alert host defences to the pathogen attack (Leulier & Lemaitre 2008). PAMPs recognised by TLRs undergo selective sweeps, and TLR genes must adapt rapidly in response – here, their variation was associated with directional selection. Note that there was a significant surplus of nonsynonymous SNPs (81% of the total) in TLR extracellular domains, which constituted just 59% of the protein sites. Variation in the extracellular domain of TLR4 has already been associated with disease (Beaumont et al. 2003, Leveque et al. 2003, Ye et al. 2006). Therefore, in response to novel pathogens encountered in newly occupied niches, chicken TLR genes have been driven by a functional requirement to adapt at sites that interact with PAMPs (Barreiro et al. 2009).

In contrast to TLRs, cytokines are mediating molecules that initiate pro-inflammatory signals in the immune system in response to parasitic, bacterial or viral infections (Kaiser et al. 2005). Despite considerable sequence divergence, avian and mammalian cytokines share many common features (Staeheli et al. 2001): in mammals, cytokines have roles regulating the expression of the innate, cell-mediated (T_H1) and humoral (T_H2) immune responses – homologous chicken cytokines are likely to possess similar functions (Kaiser et al. 2005, Avery et al. 2004). The pattern of balanced variation preventing complete fixation of nonsynonymous alleles that was more perceptible at cytokine genes may be related to their functional pleiotropy and the range of proteins with which they interact. As a consequence of being focal communicating molecules, cytokines are involved in the response to many infectious diseases and thus their selection signatures may be more variegated. Sharp adaptive sweeps at cytokine genes may compromise immunity to certain pathogens, so populations that thrive could be those where functional diversity remains high, a form of frequency-dependent selection (Asthana et al. 2005). Sustained periods of positive selection would also cause a loss of heterozygosity: this was more appreciable for TLRs than cytokines.

Furthermore, cytokines fine-tune the immune response by regulating its scope, and so their roles may be improved by the addition of functions and cell targets (Ferrer-Admetlla et al. 2008). There could be selective pressure to regulate cytokine gene expression (Beaty et al. 1995, Williams et al. 2000, Li et al 2009) – a phenomenon observed at the chicken MHC-B locus (O'Neill et al. 2009). In addition, the requirement of synergistic adaptation at cytokines and their target receptors may limit the viability of polymorphisms in active proteins of chicken cytokines.

Cytokines also have a wide range of physiological effects and have been shown to influence thermoregulation, embryonogenesis, appetite, apoptosis, gut motility and vascular endothelium activation (Conti et al. 2004, Pfeffer 2003, Zhang 2008). Altered cytokine production or activity is likely therefore to have more significant pathological relevance than just its possible impact on resistance to infectious disease. The extensive variation at cytokine genes could therefore have been preserved in order to regulate the balance between responding effectively to infectious diseases and maintaining normal biological activity (Ferreira 2003, Ferrer-Admetlla et al. 2008).

Reduced rates of low-frequency nonsynonymous mutations at cytokines could also be a consequence of conservation to counteract potential pathological implications. Excess production of cytokine inflammatory signals is associated with atopic pathologies in humans (Howard et al. 2000) and with more severe responses to infectious challenge, such as malaria (Khorr et al. 2007). Although signals initiating atopy and the response to infection are conveyed through TLRs in humans (Akira et al. 2006), their role in recognising PAMPs may mean that immune response power must be regulated at a different pathway level: signals of balancing selection may be present at immune genes regulating inflammation (Andrès et al. 2009). While an inadequate cytokine reaction could lead to death, excessive cytokine response leading to cytokine storm or allergic inflammation (Hoffjan et al. 2003) could also lead to death or at least would be energetically expensive (Ots et al. 2001), thus a pathogen-pathology equilibrium could operate within a metabolic framework (Chiang et al. 2009). This is clearly dependent on pathogen virulence and diversity, which has varied both temporally and geographically in chicken history, where an array of isolated populations may have originally adapted to different local disease challenges and later migrated extensively before amalgamating. Thus the sustained maintenance of balanced variability at cytokines could be a form of frequency-dependent selection driven by selective forces from a wide range of microorganisms because of their numerous key functional roles. This could be a general property of vertebrate cytokine genes (Wilson et al. 2006, Mege et al. 2006).

Differing patterns of diversity between gene classes have been observed elsewhere: recent surveys of human diversity at innate immune receptor, mediator and effector genes found evidence for both directional and balancing selection (Ferrer-Admetlla et al. 2008, Wilson et al. 2006). A one-tailed t-test of the mean Tajima's *D* of 68 cytokine and cytokine receptor genes and of 238 human NIEHS genes for Europeans showed that the cytokine *D* value was significantly higher (0.41 ± 0.89 for cytokines vs -0.04 ± 1.03, p = 7.02 x 10^{-4}; Fumagalli et al. 2009), and was supported by tests of other data (0.40 ± 0.76 for 30 cytokines vs 0.09 ± 0.92 for 102 other genes, p = 0.049; Akey et al. 2004). However, neither datasets showed a significant difference for African-American samples, which may be a consequence of their lower DAFs (Fredman et al. 2006).

The selective processes acting on plant immune genes differ according to the functional categories of their gene products (Moeller & Tiffin 2008). The pattern of stronger directional selection at receptor than at mediator immune genes has also been observed in *Drosophila*, and may be related to the gene products' positions in immune signalling networks, which reflect their functional roles (Sackton et al. 2007). Components on the periphery of such networks, like receptors, that interact with a defined set of adapting pathogen molecules can adapt more specifically than those located more centrally and with multiple functions (Cui et al. 2009). Cellular protein networks of the probability of positive selection indicate receptors are topologically more predisposed to this than mediators, as are sites on the protein surface than those located internally – as a result, the surface area of the protein selectively constrained could increase as number of interacting partners rises (Kim et al. 2007). Therefore, the evolutionary patterns at cytokines could be restricted by their functions in binding to many receptor types, including those from beyond the immune system (Bezbradica & Medzhitov 2009), and so cytokines play a role in many biological processes. By contrast, the roles of TLRs in recognising PAMPs from a smaller number of pathogens means their domains that recognise microbial molecules can evolve more freely and more swiftly.

7.5 Conclusion

This Chapter implemented new sequencing technology to survey an array of cytokine and TLR genes, along with MC1R, in a set of chickens and red, grey, Ceylon and green JF. Solexa-based resequencing proved to be a powerful method for delivering high-coverage at key variable sites. As a result of this increase in sequencing data, enhanced population genetic tests that can analyse groups of genes rather than just single genes will need to be determined (for example, see Barreiro et al. 2009).

Analysis of MC1R showed that two amino acid-altering mutations that change plumage colour define chicken from JF, and may have been subject to selection or founding effects during chicken domestication.

Variation at the two genes classes possessed different shared and contrasting characteristics. All showed a pattern of elevated allele diversity, with no separation of chicken from red JF. This clearly indicated that red JF was the main genetic source for chicken, and illustrated that there has been continuous historical genetic exchange between the groups. At two genes, there was evidence of introgression of grey JF into the chicken, or perhaps the reverse: further wide sampling of grey and red JF is needed in order to verify this hypothesis. There was evidence of shared patterns of diversity at certain cytokine (IL13 and IL4) and TLR (1LA and 1LB) genes, consistent with the idea that adjacent immune genes may be co-expressed in response to infection.

There was a significant difference in statistics and spectra incorporating allele frequencies that affected the distribution of functional replacement substitutions between the cytokines and TLRs. These disparities did not appear to be caused by the demographic history of the chicken and its domestication. Such differences may be related to the functional roles of both groups. TLRs directly interact with pathogen molecules, and so might be subject to sharper selective processes that would likely increase the frequency of one effective allele, thus giving a signature of positive selection. Cytokines have multiple roles in mediating signals in the immune system to many microorganisms and beyond immunity as well, and thus their patterns of diversity are indicative of frequency-dependent selection. Consequently, variation at cytokines may favour several intermediate-frequency alleles, giving a signal of balancing selection. In this way, the adaptive forces regulating both gene classes are determined to a large extent by the functional roles undertaken by each group.

Publication

This chapter formed the basis for a manuscript submitted to the Journal of Immunology entitled "The differential evolutionary dynamics of chicken cytokine and toll-like receptor gene classes". The authors are: Downing T, Lloyd AT, O'Farrelly C and Bradley DG. Work on MC1R is being formatted for a paper entitled "Variation affecting plumage colour at the MC1R locus differentiates chicken and jungle fowl" by Downing T, Lloyd AT, O'Farrelly C and Bradley DG.

CHAPTER 8

Conclusion

8.1 The evolution of chicken immune genes

Chicken immune genes are subject to selection: this was demonstrated by comparing them to their zebra finch orthologs in Chapter 2. This approach allowed confident exclusion of a relaxation of selective constraint perceived in other domestic organisms, like the dog (Cruz et al. 2008), and indicated that the sequences of immune genes were both conserved on the chicken-zebra finch lineage and within chicken itself. This framework was important for determining the relevance of replacement mutations detected among chickens: alleles at high-frequencies appear likely to retain functionality (Sabeti et al. 2006), even if they have separate origins and later were mixed. In addition, genes that have evolved rapidly since the divergence of the ancestors of chicken and zebra finch, appear to retain these properties when resequenced in chicken populations, such as IL4RA. This suggests that genes subject to selective processes in avian lineages may continue to be in extant birds.

8.1.1 Identifying functional immune gene variation:

Assesing the weight of evidence for selection at each chicken gene in light of the results of the wider dataset is instructive. While the suggestion of purifying selection at coding regions of IFNG appears to be accurate, given the high levels of coding sequence variation at other genes, the proposal of balancing selection at IL1B could be a misinterpretation of the chicken population's history. This is dependent on the truth of the assertion that cytokines wre subject to frequency-dependent selective processes. At lysozyme and IL4RA, it may be that much variation first generated by the chicken's demographic history was being maintained. Given the crucial roles of both in the immune response, it is possible that this extensive diversity was driven by selective pressure to enhance interaction with other evolving molecules, while retaining those functions already present. The diversity trend at lysozyme (like IL4RA) may be caused by directional selection in separate environments and subsequent admixture; however, only measuring the enzymatic capacity of each variant can determine if this is neutral or functional. Like IFNG and lysozyme, GMCSF, IL3, IL8 and TLR7 all showed evidence of strong conservation at coding sites. Most cytokines (IL1B, GMCSF, IL12A, IL13, IL3, IL4, KK34) had balanced patterns of diversity; this contrasted with the TLRs (TLR1LA, TLR1LB, TLR2A, TLR3, TLR5, TLR7), which displayed variation more consistent with directional selection.

Highly variable geographically unstructured allele clusters appeared to be the norm at chicken genes and were sufficient to explain much of the variation present. Initially, this may suggest that genes with low levels of variation, such as TLR2B, are candidates for selection where their extreme homozygosity may reflect a recent selective sweep. However, TLR4 had high diversity segregated into nodes partitioned by disease-associated nucleotide polymorphisms, suggesting that selective processes may be obscured by demography. If different TLR4 variants were adaptively advantageous in briefly allopatric populations that later combined, the outcome in modern chickens could be similar to that observed at this gene. It is a feature of variation at TLR4 that all these disease-associated SNPs were segregating at high frequencies; challenging chickens variable at the nonsynonymous sites identified for other genes may reveal further sites of relevance to infectious disease.

8.2 The high variability of chickens

The underlying theme of chicken population genetics was a high level of intraspecific diversity, which was at least two times greater than in cattle, the only other livestock organism to have its genome sequenced (Bovine HapMap Consortium 2009), and was much more variable than humans, dogs and gorillas, but was of the same scale as rodents (International Chicken Genome Sequencing Consortium 2004). This elevated variation has been observed at many chicken genes, microsatellites and mtDNA (Liu et al. 2007, Worley et al. 2008, Berlin et al. 2008, Muchadeyi et al. 2008, Berthouly et al. 2009, Kanginakudru et al. 2008, Bao et al. 2008, Granevitze et al. 2009, Oka et al. 2007), and was present at the 23 genes studied in this thesis, with the exceptions of TLR2B and MC1R.

8.2.1 Multiple origins and hybrid histories:

Modern chickens are the product of multiple domestications of geographically distinct red JF in south, east and south-east Asia (Fumihito et al. 1996, Liu et al. 2007, Oka et al. 2007, Kanginakudru et al. 2008): the exact number of domestications is difficult to determine because of the complexity of the chicken's demographic past. It is clear that red JF was the primary contributor to genetic diversity (Fumihito et al. 1994, International Chicken Genome Sequencing Consortium 2004): this could be an artefact of its wide geographic range in comparison to the more restricted grey, Ceylon and green JF (Eriksson et al. 2008). Ceylon JF exists largely within Sri Lanka, green JF are commonly found on islands off the south-east Asian peninsula, and grey JF inhabits west and south India (West & Zhou 1989). Since only grey JF outside of red JF has been proven to has made a genetic contribution to domestic chickens (Nishibori et al. 2005, Silva et al. 2008), geographical barriers may have genetically isolated Ceylon and green JF from chicken – these are perhaps examples of allopatric speciation. Eriksson et al. (2008) showed how grey JF contributed the trait of yellow legs to domestic chickens: more genes may follow this pattern, though perhaps with more subtle effects. Here, at TLR1LA and TLR2A there is evidence of grey JF introgression; however, further resequencing of grey and red JF as well as global chicken populations are necessary to determine this fully. Additionally, by assaying the immune response to infections of birds with known TLR1LA and TLR2A genotypes variable between and within chicken and grey JF, it would be possible to determine the functional consequences of this shared variation.

8.2.2 Admixture of disparate chicken populations:

Additional to the processes of domestication and introgression, the chicken's considerable diversity is likely to have been increased by the subsequent admixture of previously separate groups. The chicken's physical attributes of being a portable egg-laying and meat-bearing domestic animal historically may have lent itself to widespread global dispersal driven by humans engaged in migration and trade (Berthouly et al. 2009 Muchadeyi et al. 2008). In the networks of IL1B, IFNG, IL4RA and lysozyme, not only was every continental group represented in each major allele cluster, but every population was as well. The transcontinental human-driven spread of chickens resulted in different populations with very high variability and origins from geographically distant regions (Razafindraibe et al. 2009, Kanginakudru et al. 2008, Berthouly et al. 2009, Granevitze et al. 2009, Muchadeyi et al. 2008). As a result, the Asian, African, heritage and broiler chickens when resequenced all displayed high allelic diversity and little population structure. The absence of a genetic link with geography is unusual among edible domestic animals (Bruford et al. 2003), and is a reflection of how the chicken's complex demographic past has defined its current level of genomic variation. Moreover, it is possible the chicken's elevated diversity historically enhanced its ability to spread to new geographic niches (Merilä 2009).

The chicken's traits of portability, a short generation time, a capacity to breed with related species, and an ability to produce multiple food products for humans caused a significant component of its molecular genomic diversity. Its pattern of high variation originating in an indefinite but undoubtedly large number of domestications contrasts strongly with livestock that were only domesticated a limited number of times, including cattle (Troy et al. 2001), pigs (Bruford et al. 2003), sheep (Hiendleder et al. 1998) – and possibly horses and goats (Bruford et al. 2003). Other mobile domestic species, such as dogs, have migrated extensively, but may only have undergone as few as one or two domestications (Boyko et al. 2009, Pang et al. 2009). Chicken may share more features with other hybrid species, such as cereal crops, which have multiple origins and thus show extensive levels of variability (Salamini et al. 2002).

8.3 Factors defining chicken immune diversity

The chicken's genetic history has generated great diversity which continues to be maintained in extant birds. Given that chicken immune defences are attacked by a wide range of pathogens, it is essential to possess an effective response to these onslaughts. Therefore there may be an adaptive advantage for chicken populations in preserving immune system variability when different sets of novel local microbes attempt to infect them. This impetus to conserve variation is more acute in chicken than in other animals because of avian genomic reductions in gene family size, but is somewhat ameliorated by the wide immune repertoire of molecular as well as gene family diversity (Kaiser et al. 2005).

8.3.1 A possible heterozygosity-fitness correlation:

The pool of genetic diversity created by the chicken's demographic history may serve as an immunological arsenal against pathogens. A heterozygosity-fitness correlation could exist in chickens similar to that in humans (Lyons et al. 2009a, Lyons et al. 2000b), where extensive homozygosity at key immune genes reduces the ability to fight off infectious diseases. A sexual preference for partners with divergent genotypes is present in song sparrows (*Melospiza melodia*; Reid et al. 2007) and in black-legged kittiwakes (*Rissa tridactyla*; Mulard et al. 2009). Consequently, diversity generated by demography and sexual selection may be effective in increasing the genetic fitness of the population (Tiemann & Rehkämper 2009). This has been classically asserted for the MHC in chicken (Worley et al. 2008) and for other organisms (see Jeffery & Bangham 2000, Bernatchez & Landry 2003, Wu et al. 2001, Hoffjan et al. 2003). It is also supported by high levels of recombinant variants at chicken immune genes, which both contribute further to diversity (Charlesworth 2006) and reduce a loss of neutral diversity associated with linkage to deleterious alleles (Hill & Robertson 1966). BGC may play a role in generating diversity: the average GC content for all 23 resequenced genes is higher at the 3rd codon position than that for the CDS as a whole (0.58 ± 0.14 vs 0.49 ± 0.12 for CDS, t-test $p = 0.024$). Since the 3rd codon position is more frequently redundant, this suggests variants with G and C nucleotides are fixed more frequently than those with A or T, a sign of BGC.

8.3.2 Mitigating pathogenicity, atopy and metabolism:

Homozygosity and heterozygosity at immune genes may be both advantageous and disadvantageous for a bird in an environment with an array of adapting microorganisms with varying pathogenicity. A heterozygous bird is more likely to be react effectively to the evolution of a novel variant than a homozygote. However, if an infection requires a strong immune response, a homozygous bird may be favoured if resistance to infection in the heterozygote is compromised. Yet, overzealous immunity may damage the fitness of the organism by developing autoimmune diseases (Link 1998, Kay et al. 2009, Khor et al. 2007) – the presence of atopic effects may be caused by the occasional need for strong immune responses. Consequently, once the immune system has evolved to tolerate a pathogen, a form of heterozygote advantage may exist until the development of a new microbe variant, leading to a pattern of balanced diversity (Charlesworth 2006).

The maintenance of variation is not just a result of allergy-driven heterozygote advantage or pathogen-based directional selection, but also of the population-wide consequences of a balance between these two effects, in addition to the historical demographic forces described above. When moved by humans to new environments, populations may undergo selective sweeps in response to new virulent pathogens – however, the genetic fitness of such birds may be decreased by allergies. Accordingly, the persistence of multiple alleles could be selectively advantageous if the effects of excessive inflammation are sufficiently costly in a pathogen-rich environment. The detection of these effects would be amplified by latent admixture given the local nature of such directional selection events in combatting virulent diseases.

Inferring the consequences of atopy for ancient farming has relevance to modern chicken breeds, which show reduced diversity when compared to wild or village chickens (International Chicken Polymorphism Map Consortium 2004, Muir et al. 2008). There is a metabolic cost to mounting an immune response: in wintering great tits (*Parus major*) the associated loss of weight is significant (Ots et al. 2001). In addition, allergies and autoimmune diseases that mount an immune reaction where none is needed have a metabolic cost (Demas et al. 1997, Lochmiller & Deerenberg 2000, Bonneaud et al. 2003, Khor et al. 2007), which would have reduced the growth, reproductive and food production rates for domestic chickens, reducing their fitness.

Additionally, there is evidence the immune system can evolve to optimise energy usage (Råberg et al. 2002). Therefore, it is likely that diversity in ancient chicken flocks was balanced by the ability to combat disease and to avoid unnecessary allergies. Although chickens are diverse in comparison to other domesticated animals, a continued loss of diversity in commercial breeds may impact on their ability to be metabolically efficient, though a capability to outbreed with wild JF provides the possibility to offset this effect.

The metabolic costs associated with the immune response may be part of a more universal theme of metabolic imperatives shaping diversity in the avian genome. The chicken genome is three times smaller than those of mammalian species as a result of reductions in the number and length of repeats, in intron length and in gene family number (International Chicken Genome Sequencing Consortium 2004). It is suggested that the evolution of flight in birds was the main cause behind this adaptation because smaller genomes improve birds' rate of oxidative metabolism (Hughes & Hughes 1995). This change is paralleled by smaller cell sizes in birds than mammals (Hughes & Friedman 2008): genome mass and cell size correlate in vertebrates (Szarski 1976). This size reduction occurs in all gene families, including immune system ones (Hughes & Friedman 2008) and it may be related to the high rate of chromosomal rearrangements in the avian lineage (Hannson et al. 2009, Nie et al. 2009, Griffin et al. 2007, Stapley et al. 2008, Itoh & Arnold 2005).

As a result of extensive purifying selection against gene duplication and to prevent gene function redundancy, a situation of antagonistic pleiotropy may arise at immune genes, where there is no "escape from adaptive conflict" (EAC) by different alleles as they are driven by pathogens to optimise the multiple functions of the protein (Hughes 1994, Des Marais & Rausher 2008). EAC is a form of subfunctionalisation that differs from the classical neofunctionalisation MDN (mutation during non-functionality) model because new functions that do not replace current functions are acquired prior to duplication (Conant & Wolfe 2009). This may result in the pattern of balancing selection observed in certain immune genes, where the improvement of duplication and new immune gene functions would need to far outweigh the disadvantage of a slightly bigger genome. Generally, at least one in a pair of duplicated genes (like TLR1LA/B and TLR2A/B) would be expected to have to have undergone rapid sub- or neo- or sub-neo-functionalisation following the duplication event (Storz 2009). The difference between EAC subfunctionalisation and that of the DDC (duplication, degeneration, complementation) model is the variation performing distinctive subfunctions would be present prior to the duplication event (Conant & Wolfe 2009). A genome-wide interspecies analysis of orthologs and paralogs would be required to test this confidently: in *Drosophila*, the number of novel genes as well as gene gains and losses in the mediator class is significantly fewer than those in the receptor or effector groups (Sackton et al. 2007, Cormican et al. 2009). This suggests that an EAC scenario may apply more frequently to proteins like pleiotropic cytokines than to other immune categories. It may be that not only the metabolic demands of flights cause a reduction in the genome, repetitive sequence and gene family size, but the energetic costs of an effective immune response continue to induce selective process on relevant chicken genes.

8.3.3 Selection on standing variation:

Selective forces operating on chicken immune genes may act more frequently on standing variability than on new mutations. In such events, a non-singleton allele that was previously neutral becomes advantageous – these tend to initiate "soft" sweeps (Innan & Kim 2004). These selection signatures are different to those where a *de novo* variant is selected – a "hard" sweep – which can include migrant alleles (for example, wild JF; Hermisson & Pennings 2005). The difference in diversity lost between hard and soft selective sweeps is determined by the initial frequency of the selected allele (Przeworski et al. 2005). If the adaptive evolution of standing diversity is common for chicken, the frequency of the allele under selection would be more important than the strength of selection (Innan & Kim 2004). And given that the probability of fixation of an advantageous allele increases linearly with its initial frequency (Barrett & Schluter 2007), the selective pressure required to fix alleles already segregating in a population could be lower.

At chicken immune genes there is evidence that many SNPs are both under selection and yet, incongruously under a hard sweep model, variability remains high. Pathogens take many different forms and would simultaneously be stimulating reactions from and adaptation of the chicken immune system: thus rather than having one superior chicken allele sweep to fixation within a population, a concomitant immune battle with an array of infectious challenges may result in selection for a landscape of varying adaptive fitnesses (Arnold et al. 2001). This would also be dependent on the mildly deleterious effects of atopy, and would have occurred in the context of changing environments with new selective pressures. Consequently, this "arms race" may lead to previously neutral mutations becoming slightly advantageous (Pennings & Hermisson 2006a). If such softer sweeps allow the immune system to make the minimum modification possible and preserve necessary functions already present, this may temporarily optimise fitness in a superior manner compared to hard sweeps if waves of microbial challenges are sufficiently frequent. Softer sweeps at vertebrate immune genes may also help explain the modern day persistence of ancient disease-susceptible alleles at unexpectedly high frequencies (Wakeley 2008).

Under a classical hard sweep model, much variation would be lost (Barrett & Schluter 2007), and significant differentiation would be expected following the subsequent admixture of populations – this was rarely observed here. Accordingly, it is likely that a combination of hard and soft sweeps act on disease-associated genes (Pennings & Hermisson 2006b). There are instances of some population- and species-level divergence, particularly at TLR genes, indicating the presence of hard sweeps. Additionally, the TLR genes had a lower θ_W compared to the cytokines, indicating that relatively more of their adaptive novel mutations may behave like hard sweeps (Pennings & Hermisson 2006a). Therefore, genes that interact with the environment like TLRs may have more frequent hard sweeps than signalling molecules, such as cytokines, where "Red Queen" selection may be more prevalent (Hurst 2009). This discounts the occurrence of recurrent or fluctuating selection at cytokines, because this would favour rare- and high-frequency variants (Kim 2006, Huerta-Sanchez et al. 2008).

8.3.4 The functional role of the gene product:

This difference between gene classes is a result of an additional constraint on the evolutionary patterns at immune genes: their functional category. This was supported by the different patterns of allele frequency spectra at cytokine and TLR genes, and in other studies of immune gene classes in humans (Ferrer-Admetlla et al. 2008, Fumagalli et al. 2009), plants (Moeller & Tiffin 2008), and *Drosophila* (Sackton et al. 2007). The effects of protein interaction (Cui et al. 2009) and cellular networks (Kim et al. 2007) on selective processes are clear: components on the periphery, at the level of a cell or protein, evolve faster than parts located more centrally, which are more conserved because they interact with more molecules and so there is reduced protein surface area for modification. For signalling molecules like cytokines, the requirement to synergistically evolve with their cytokine receptors on the cell membrane would further modify their variability. A further facet of this limitation on adaptive novelty is the differing dynamics of protein domains that interact with pathogens, which are under stronger selection to change than domains that do not (Barreiro et al. 2009). Hence the functional role of protein domains may reflect the selective signature present (Barreiro et al. 2009), such that different regions can produce divergent signals.

In summary, a complex demographic history of multiple domestications, outbreeding with wild JF and the subsequent admixture of previously separate populations has resulted in a high level of chicken genetic diversity. Even though immune genes are subject to selection pressures from infectious diseases, this variation is maintained because of the changing consequences of immune response intensity in disparate local environments with different pathogens that adapt over time. The pattern of variability at each immune locus is further defined by not only the functional category of its product but also by the specific role of each domain within that protein.

REFERENCES

Abasht B, Kaiser MG, Lamont SJ (2008) Toll-like receptor gene expression in cecum and spleen of advanced intercross line chicks infected with *Salmonella enterica* serovar Enteritidis. *Vet Immunol Immunopathol*. 123:314-23.

Abasht B, Kaiser MG, van der Poel J, Lamont SJ (2009) Genetic lines differ in Toll-like receptor gene expression in spleens of chicks inoculated with *Salmonella enterica* serovar Enteritidis. *Poult Sci.* 88:744-9.

Abdul-Careem MF, Hunter DB, Lambourne MD, Read LR, Parvizi P, Sharif S (2008) Expression of cytokine genes following pre- and post-hatch immunization of chickens with herpesvirus of turkeys. *Vaccine* 26:2369-77.

Agarwal S, Rao A (1998) Long-range transcriptional regulation of cytokine gene expression. *Curr. Opin. Immunol.* 10:345–52

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* 19(5):711-22.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2(10):e286.

Akey JM, Zhang K, Xiong M, Jin L (2003) The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol Biol Evol.* 20(2):232-42.

Akira S, Hoshino K, Kaisho T (2000) The role of Toll-like receptors and MyD88 in innate immune responses. *J Endotoxin Res.* 6:383-7.

Akira S, Uematsu S, Takeuchi O (2006) Pathogen recognition and innate immunity. *Cell* 124(4):783-801.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389-402.

Anderle C, Hammer A, Polgár B, Hartmann M, Wintersteiger R, Blaschitz A, Dohr G, Desoye G, Szekeres-Barthó J, Sedlmayr P (2008) Human trophoblast cells express the immunomodulator progesterone-induced blocking factor. *J Reprod Immunol.* 79(1):26-36.

Anderson TM, vonHoldt BM, Candille SI, Musiani M, Greco C, Stahler DR, Smith DW, Padhukasahasram B, Randi E, Leonard JA, Bustamante CD, Ostrander EA, Tang H, Wayne RK, Barsh GS (2009) Molecular and evolutionary history of melanism in North American gray wolves. *Science* 323(5919):1339-43.

Andersson L (2003) Melanocortin receptor variants with phenotypic effects in horse, pig, and chicken. *Ann N Y Acad Sci.* 994:313-8.

Andolfatto P (2008) Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics* 180(3):1767-71.

Andreozzi L, Federico C, Motta S, Saccone S, Sazanova AL, Sazanov AA, Smirnov AF, Galkina SA, Lukina NA, Rodionov AV, Carels N, Bernardi G: (2001) Compositional mapping of chicken chromosomes and identification of the gene-richest regions. *Chromosome Res.* 9:521-32.

Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, Clark AG, Nielsen R (2009) Targets of balancing selection in the human genome. *Mol Biol Evol.* Epub Aug 27th.

Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18: 1585–1592.

Anisimova M, Liberles DA (2007) The quest for natural selection in the age of comparative genomics. *Heredity* 99(6):567-79.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 32(Database issue):D115-9.

Ardell, DH (2004) SCANMS: adjusting for multiple comparisons in sliding window neutrality tests. *Bioinformatics* 20(12):1986-8.

Ask B, van der Waaij EH, Glass EJ, Bishop SC. (2007) Modelling immunocompetence development and immunoresponsiveness to challenge in chicks. *Poultry Science* 86:1336-50.

Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S (2007) Analysis of sequence conservation at nucleotide resolution. *PLoS Comput Biol.* 3(12):e254.

Asthana S, Schmidt S, Sunyaev S (2005) A limited role for balancing selection. *Trends Genet.* 21(1):30-2.

Avery S, Rothwell L, Degen WD, Schijns VE, Young J, Kaufman J, Kaiser P (2004) Characterization of the first nonmammalian T2 cytokine gene cluster: the cluster contains functional single-copy genes for IL-3, IL-4, IL-13, and GM-CSF, a gene for IL-5 that appears to be a pseudogene, and a gene encoding another cytokinelike transcript, KK34. *J Interferon Cytokine Res*. 24:600-10.

Axelsson E, Ellegren H. 2009. Quantification of adaptive evolution of genes expressed in avian brain and the population size effect on the efficacy of selection. Mol Biol Evol. 26:1073-9.

Axelsson E, Hultin-Rosenberg L, Brandström M, Zwahlén M, Clayton DF, Ellegren H (2008) Natural selection in avian protein-coding genes expressed in brain. *Mol Ecol*. 17(12):3008-17.

Axelsson E, Webster MT, Smith NG, Burt DW, Ellegren H (2005) Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Res*. 15(1):120-5.

Baguet A, Sun X, Arroll T, Krumm A, Bix M (2005) Intergenic transcription is not required in Th2 cells to maintain histone acetylation and transcriptional permissiveness at the Il4-Il13 locus. *J Immunol*. 175(12):8146-53.

Baião PC, Schreiber E, Parker PG (2007) The genetic basis of the plumage polymorphism in red-footed boobies (*Sula sula*): a melanocortin-1 receptor (MC1R) analysis. *J Hered*. 98(4):287-92.

Balkissoon D, Staines K, McCauley J, Wood J, Young J, Kaufman J, Butter C (2007) Low frequency of the Mx allele for viral resistance predates recent intensive selection in domestic chickens. *Immunogenetics* 59(8):687-91.

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. *Nat Rev Genet*. 4(2):99-111.

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16:37-48.

Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics* 141(2):743-53.

Bao W, Chen G, Li B, Wu X, Shu J, Wu S, Xu Q, Weigend S (2008) Analysis of genetic diversity and phylogenetic relationships among red jungle fowls and Chinese domestic fowls. *Sci China C Life Sci*. 51(6):560-8.

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B, Kidd JR, Kidd KK, Alcaïs A, Ragimbeau J, Pellegrini S, Abel L, Casanova JL, Quintana-Murci L (2009) Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet*. 5(7):e1000562.

Barrett RD, Schluter D (2008) Adaptation from standing genetic variation. *Trends Ecol Evol*. 23(1):38-44.

Beaty JS, West KA, Nepom GT (1995) Functional effects of a natural polymorphism in the transcriptional regulatory sequence of HLA-DQB1. *Mol Cell Biol.* 15(9):4771-82.

Beaumont C, Protais J, Pitel F, Leveque G, Malo D, Lantier F, Plisson-Petit F, Colin P, Protais M, Le Roy P, Elsen JM, Milan D, Lantier I, Neau A, Salvat G, Vignal A (2003) Effect of two candidate genes on the Salmonella carrier state in fowl. *Poult Sci.* 82:721-6.

Bedard PA, Alcorta D, Simmons DL, Luk KC, Erikson RL (1987) Constitutive expression of a gene encoding a polypeptide homologous to biologically active human platelet protein in Rous sarcoma virus-transformed fibroblasts. *Proc Natl Acad Sci U S A.* 84:6715-9.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* 340(4):783-95.

Bergin AM, Balder B, Kishore S, Swärd K, Hahn-Zoric M, Löwhagen O, Hanson LA, Padyukov L (2006) Common variations in the IL4R gene affect splicing and influence natural expression of the soluble isoform. *Hum Mutat.* 27(10):990-8.

Berglund J, Pollard KS, Webster MT (2009) Hotspots of biased nucleotide substitutions in human genes. *Plos Biol.* 7(1):e26.

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J Evol Biol.* 16(3):363-77

Berlin S, Qu L, Li X, Yang N, Ellegren H (2008) Positive diversifying selection in avian Mx genes. *Immunogenetics* 60(11):689-97.

Berndt A, Wilhelm A, Jugert C, Pieper J, Sachse K, Methner U (2007) Chicken cecum immune response to *Salmonella enterica* serovars of different levels of invasiveness. *Infect Immun.* 75:5993-6007.

Berthouly C, Leroy G, Van TN, Thanh HH, Bed'Hom B, Nguyen BT, Vu CC, Monicat F, Tixier-Boichard M, Verrier E, Maillard JC, Rognon X (2009) Genetic analysis of local Vietnamese chickens provides evidence of gene flow from wild to domestic populations. *BMC Genet.* 10:1.

Betrán E, Rozas J, Navarro A, Barbadilla A (1997) The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. *Genetics* 146(1):89-99.

Bezbradica JS, Medzhitov R (2009) Integration of cytokine and heterologous receptor signaling pathways. *Nat Immunol.* 10(4):333-9.

Blake CC, Koenig DF, Mair GA, North AC, Phillips DC, Sarma VR (1965). Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. *Nature* 206(986):757-61.

Boardman PE, Sanz-Ezquerro J, Overton IM, Burt DW, Bosch E, Fong WT, Tickle C, Brown WR, Wilson SA, Hubbard SJ (2002) A comprehensive collection of chicken cDNAs. *Curr Biol.* 12(22):1965-9.

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST--database for "expressed sequence tags". *Nat Genet.* 4(4):332-3.

Bonifer C, Jägle U, Huber MC (1997) The chicken lysozyme locus as a paradigm for the complex developmental regulation of eukaryotic gene loci. *J Biol Chem.* 272(42):26075-8.

Bonneaud C, Mazuc J, Gonzalez G, Haussy C, Chastel O, Faivre B, Sorci G (2003) Assessing the cost of mounting an immune response. *Am Nat.* 161(3):367-79.

Bovine HapMap Consortium (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324(5926):528-32.

Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhano M, Corey L, Degenhardt J, Auton A, Hedimbi M, Kityo R, Ostrander EA, Schoenebeck J, Todhunter RJ, Jones P, Bustamante CD (2009) Complex population structure in African village dogs and its implications for inferring dog domestication history. *Proc Natl Acad Sci U S* A. Epub Aug 3rd.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4(5):e1000083.

Bratt J, Palmblad J (1997) Cytokine-induced neutrophil-mediated injury of human endothelial cells. *J Immunol.* 159(2):912-8.

Bruford MW, Bradley DG, Luikart G (2003) DNA markers reveal the complexity of livestock domestication. *Nat Rev Genet.* 4(11):900-10.

Buard J, de Massy B (2007) Playing hide and seek with mammalian meiotic crossover hotspots. *Trends Genet.* 23(6):301-9.

Burt DW (2005) Chicken genome: current status and future opportunities. *Genome Res.* 15(12):1692-8.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153-7.

Caldwell RB, Kierzek AM, Arakawa H, Bezzubov Y, Zaim J, Fiedler P, Kutter S, Blagodatski A, Kostovska D, Koter M, Plachy J, Carninci P, Hayashizaki

Y, Buerstedde JM (2005) Full-length cDNAs from chicken bursal lymphocytes to facilitate gene function analysis. *Genome Biol.* 6(1):R6.

Candille SI, Kaelin CB, Cattanach BM, Yu B, Thompson DA, Nix MA, Kerns JA, Schmutz SM, Millhauser GL, Barsh GS (2007) A β-defensin mutation causes black coat color in domestic dogs. *Science* 318(5855):1418-23.

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA (2005) Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15(11):1553-65.

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet.* 7(2):98-108.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2(4):e64.

Charo IF, Ransohoff RM (2006) The many roles of chemokines and chemokine receptors in inflammation. *N Engl J Med.* 354(6):610-21.

Charpentier MJ, Widdig A, Alberts SC (2007) Inbreeding depression in non-human primates: a historical review of methods used and empirical data. *Am J Primatol.* 69(12):1370-86.

Cheviron ZA, Hackett SJ, Brumfield RT (2006) Sequence variation in the coding region of the melanocortin-1 receptor gene (MC1R) is not associated with plumage variation in the blue-crowned manakin (*Lepidothrix coronata*). *Proc Biol Sci.* 273(1594):1613-8.

Chiang HI, Swaggerty CL, Kogut MH, Dowd SE, Li X, Pevzner IY, Zhou H (2008) Gene expression profiling in chicken heterophils with Salmonella enteritidis stimulation using a chicken 44 K Agilent microarray. *BMC Genomics* 9:526.

Chiang SH, Bazuine M, Lumeng CN, Geletka LM, Mowers J, White NM, Ma JT, Zhou J, Qi N, Westcott D, Delproposto JB, Blackwell TS, Yull FE, Saltiel AR (2009) The Protein Kinase IKKε Regulates Energy Balance in Obese Mice. *Cell* 138:961-975.

Chubb AL (2004) New nuclear evidence for the oldest divergence among neognath birds: the phylogenetic utility of ZENK (i). *Mol Phylogenet Evol.* 30(1):140-51.

Chung CD, Liao J, Liu B, Rao X, Jay P, Berta P, Shuai K (1997) Specific inhibition of Stat3 signal transduction by PIAS3. *Science* 278(5344):1803-5.

Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. *Bioinformatics*. 20(3):426-7.

Coffman RL, Ohara J, Bond MW, Carty J, Zlotnik A, Paul WE (1986) B cell stimulatory factor-1 enhances the IgE response of lipopolysaccharide-activated B cells. *J Immunol*. 136(12):4538-41.

Conant GC (2009) Neutral evolution on mammalian protein surfaces. *Trends Genet*. 25(9):377-81.

Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet*. 9(12):938-50.

Cormican P, Lloyd AT, Downing T, Connell S, Bradley DG, O'Farrelly C (2009) The Avian Toll-Like-Receptor Pathway – subtle differences amidst general conformity. *Developmental & Comparative Immunology* 33(9):967-73.

Costantini M, Di Filippo M, Auletta F, Bernardi G (2007) Isochore pattern and gene distribution in the chicken genome. *Gene* 400:9-15.

Cox A. ELAND: Efficient Local Alignment of Nucleotide Data. (unpublished).

Cruz F, Vilà C, Webster MT (2008) The legacy of domestication: accumulation of deleterious mutations in the dog genome. *Mol Biol Evol*., 25(11):2331-6.

Cui Q, Purisima EO, Wang E (2009) Protein evolution on a human signaling network. *BMC Syst Biol*. 3:21.

Das S, Mohamedy U, Hirano M, Nei M, Nikolaidis N (2009) Analysis of the immunoglobulin light chain genes in zebra finch: evolutionary implications. *Mol Biol Evol*. Epub Sept 10th.

Davison TF (2003) The immunologists' debt to the chicken. *Br Poult Sci*. 44(1):6-21.

Dawson DA, Akesson M, Burke T, Pemberton JM, Slate J, Hansson B (2007) Gene order and recombination rate in homologous chromosome regions of the chicken and a passerine bird. *Mol Biol Evol*. 24(7):1537-52.

De Nardo D, Masendycz P, Ho S, Cross M, Fleetwood AJ, Reynolds EC, Hamilton JA, Scholz GM (2005) A central role for the Hsp90.Cdc37 molecular chaperone module in interleukin-1 receptor-associated-kinase-dependent signaling by toll-like receptors. *J Biol Chem*. 280(11):9813-22.

De S, Lopez-Bigas N, Teichmann SA (2008) Patterns of evolutionary constraints on genes in humans. *BMC Evol Biol*. 8:275.

Deichmann K, Bardutzky J, Forster J, Heinzmann A, Kuehr J (1997)Common polymorphisms in the coding part of the IL4-receptor gene. Biochem *Biophys Res Commun*. 231(3):696-7.

Delany ME (2004) Genetic variants for chick biology research: from breeds to mutants. *Mech Dev*. 121(9):1169-77.

Demas GE, Chefer V, Talan MI, Nelson RJ (1997) Metabolic costs of mounting an antigen-stimulated immune response in adult and aged C57BL/6J mice. *Am J Physiol.* 273(5 Pt 2):R1631-7.

Deng WD, Shu W, Yang SL, Shi XW, Mao HM (2009) Pigmentation in Black-boned sheep (*Ovis aries*): association with polymorphism of the MC1R gene. *Mol Biol Rep.* 36(3):431-6.

Depaulis F, Mousset S, Veuille M (2001) Haplotype tests using coalescent simulations conditional on the number of segregating sites. *Mol Biol Evol.* 18(6):1136-8.

Depaulis F., Veuille M. (1998) Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution* 15, 1788-90.

Des Marais DL, Rausher MD (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454(7205):762-5.

Dhiman N, Ovsyannikova IG, Vierkant RA, Pankratz VS, Jacobson RM, Poland GA (2008) Associations between cytokine/cytokine receptor single nucleotide polymorphisms and humoral immunity to measles, mumps and rubella in a Somali population. *Tissue Antigens* 72(3):211-20.

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature* 418(6898):700-7.

Dil N, Qureshi MA (2002) Involvement of lipopolysaccharide related receptors and nuclear factor kappa B in differential expression of inducible nitric oxide synthase in chicken macrophages from different genetic backgrounds. *Vet Immunol Immunopathol.* 88:149-61.

Downing T, Lynn DJ, Connell S, Lloyd AT, Bhuiyan AK, Silva P, Naqvi AN, Sanfo R, Sow RS, Podisi B, O'Farrelly C, Hanotte O, Bradley DG (2009a) Contrasting evolution of diversity at two disease-associated chicken genes. *Immunogenetics* 61:303-14.

Downing T, Lynn DJ, Connell S, Lloyd AT, Bhuiyan AK, Silva P, Naqvi A, Sanfo R, Sow RS, Podisi B, O’Farrelly C, Hanotte O, Bradley DG (2009b) Bioinformatic discovery and population-level validation of selection at the chicken interleukin-4 receptor alpha-chain gene. *BMC Evolutionary Biology* 9:136.

Downing T, O’Farrelly C, Bhuiyan AK, Silva P, Naqvi A, Sanfo R, Sow RS, Podisi B, Hanotte O, Bradley D (2009c) Evidence of balancing selection affecting catalytic sites of chicken lysozyme. *Animal Genetics* (in press).

Downing T, Cormican P, O’Farrelly C, Bradley DG, Lloyd AT (2009d) Evidence of the adaptive evolution of immune genes in chicken. *BMC Research Notes* (under review).

Downing T, Lloyd AT, O'Farrelly C, Bradley D (2009e) The differential evolutionary dynamics of chicken cytokine and toll-like receptor gene classes. *Journal of Immunology* (under review).

Ducrest AL, Keller L, Roulin A (2008) Pleiotropy in the melanocortin system, coloration and behavioural syndromes. *Trends Ecol Evol*. 23(9):502-10.

Duret L, Arndt PF (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 9;4(5):e1000071.

Eldon B, Wakeley J (2009) Coalescence times and FST under a skewed offspring distribution among individuals in a population. *Genetics* 181(2):615-29.

Ellegren H (2005) The avian genome uncovered. *Trends Ecol Evol*. 20(4):180-6.

Ellegren H (2007) Molecular evolutionary genomics of birds. *Cytogenet Genome Res*. 117:120-30.

Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. *Mol Ecol*. 17(21):4586-96.

Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Strömstedt L, Wright D, Jungerius A, Vereijken A, Randi E, Jensen P, Andersson L (2008) Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. *PLoS Genet*. 4(2):e1000010.

Ewers C, Antão EM, Diehl I, Philipp HC, Wieler LH. (2009) Intestine and environment of the chicken as reservoirs for extraintestinal pathogenic Escherichia coli strains with zoonotic potential. *Appl Environ Microbiol*. 75(1):184-92.

Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theor. Pop. Biol*. 3:87-112.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred II. Error probabilities. *Genome Res*. 8(3):186-94.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 8(3):175-85.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131(2):479-91.

Eyre-Walker A (2002) Changing effective population size and the McDonald-Kreitman test. *Genetics* 162(4):2017-24.

Eyre-Walker A (2006) The genomic rate of adaptive evolution. *Trends Ecol Evol*. 21(10):569-75.

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet.* 8(8):610-8.

Fang M, Larson G, Ribeiro HS, Li N, Andersson L (2009) Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS Genet.* 5(1):e1000341.

Farnell MB, Crippen TL, He H, Swaggerty CL, Kogut MH (2003) Oxidative burst mediated by toll like receptors (TLR) and CD14 on avian heterophils stimulated with bacterial toll agonists. *Dev Comp Immunol.* 27:423-9.

Faure V, Wenner T, Cooley C, Bourke E, Farr CJ, Takeda S, Morrison CG (2008) Ku70 prevents genome instability resulting from heterozygosity of the telomerase RNA component in a vertebrate tumour line. D*NA Repair (Amst).* 7(5):713-24.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155(3):1405-13.

Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. *Genetics* 158(3):1227-34.

Ferreira MA. 2003. Cytokine expression in allergic inflammation: systematic review of in vivo challenge studies. *Mediators Inflamm.* 12(5):259-67.

Ferrer-Admetlla A, Bosch E, Sikora M, Marquès-Bonet T, Ramírez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, Casals F (2008) Balancing selection is the main force shaping the evolution of innate immunity genes. *J Immunol.* 181:1315-22.

Ferrer-Costa C, Gelpi J, Zamakola L, Parraga I, de la Cruz X, Orozco M (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 21: 3176–8.

Fleming A (1922) On a remarkable bacteriolytic element found in tissues and secretions. *Proc Roy Soc Ser B* 93(653):306-317.

Fontanesi L, Tazzoli M, Beretti F, Russo V (2006) Mutations in the melanocortin 1 receptor (MC1R) gene are associated with coat colours in the domestic rabbit (*Oryctolagus cuniculus*). *Anim Genet.* 37(5):489-93.

Franjkovic I, Gessner A, König I, Kissel K, Bohnert A, Hartung A, Ohly A, Ziegler A, Hackstein H, Bein G (2005) Effects of common atopy-associated amino acid substitutions in the IL-4 receptor alpha chain on IL-4 induced phenotypes. *Immunogenetics* 56(11):808-17.

Fredman D, Sawyer SL, Strömqvist L, Mottagui-Tabar S, Kidd KK, Wahlestedt C, Chanock SJ, Brookes AJ (2006) Nonsynonymous SNPs: validation characteristics, derived allele frequency patterns, and suggestive evidence for natural selection. *Hum Mutat.* 27(2):173-86.

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147(2):915-25.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133(3):693-709.

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147:915-25.

Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Riva S, Clerici M, Bresolin N, Sironi M (2009) Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. *J Exp Med.* 206(6):1395-408.

Fumihito A, Miyake T, Sumi S, Takada M, Ohno S, Kondo N (1996) One subspecies of the red junglefowl (*Gallus gallus gallus*) suffices as the matriarchic ancestor of all domestic breeds. *Proc Natl Acad Sci U S A.* 91(26):12505-9.

Fumihito A, Miyake T, Takada M, Shingu R, Endo T, Gojobori T, Kondo N, Ohno S (1994) Monophyletic origin and unique dispersal patterns of domestic fowls. *Proc Natl Acad Sci U S A.* 93(13):6792-5.

Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23(6):273-7.

Gao F, Zhang CT (2006) Isochore structures in the chicken genome. *FEBS J.* 273(8):1637-48.

Gasparini J, Bize P, Piault R, Wakamatsu K, Blount JD, Ducrest AL, Roulin A (2009) Strength and cost of an induced immune response are associated with a heritable melanin-based colour trait in female tawny owls. *J Anim Ecol.* 78(3):608-16.

Gish W, DJ States (1993) Identification of protein coding regions by database similarity search. *Nat Genet.* 1993, 3:266-72.

Glick AD, Ranhand JM, Cole RM (1972) Degradation of group A streptococcal cell walls by egg-white lysozyme and human lysosomal enzymes. *Infection and Immunity* 6:403-13.

Göbel TW, Fluri M (1997) Identification and analysis of the chicken CD3epsilon gene. *Eur J Immunol.* 27(1):194-8.

Goetschy JF, Zeller H, Content J, Horisberger MA (1989) Regulation of the interferon-inducible IFI-78K gene, the human equivalent of the murine Mx gene, by interferons, double-stranded RNA, certain cytokines, and viruses. *J Virol.* 63(6):2616-22.

Goldstein G, Flory K, Browne B, Majid S, Ichida JM, Burtt JEH (2004) Bacterial degradation of black and white feathers. A*uk* 121:656–659.

Gongora J, Rawlence NJ, Mobegi VA, Jianlin H, Alcalde JA, Matus JT, Hanotte O, Moran C, Austin JJ, Ulm S, Anderson AJ, Larson G, Cooper A (2008) Indo-European and Asian origins for Chilean and Pacific chickens revealed by mtDNA. *Proc Natl Acad Sci U S A.* 105(30):10308-13.

Gorbach DM, Hu ZL, Du ZQ, Rothschild MF (2009) SNP discovery in Litopenaeus vannamei with a new computational pipeline. *Anim Genet.* 40(1):106-9.

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res.* 8(3):195-202.

Gordon L, Yang S, Tran-Gyamfi M, Baggott D, Christensen M, Hamilton A, Crooijmans R, Groenen M, Lucas S, Ovcharenko I, Stubbs L (2007) Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. *Genome Res.* 17(11):1603-13.

Goto T, Ohkuri T, Shioi S, Abe Y, Imoto T, Ueda T (2008) Crystal structures of K33 mutant hen lysozymes with enhanced activities. *J Biochem.*144(5):619-23.

Guo P, Hirano M, Herrin BR, Li J, Yu C, Sadlonova A, Cooper MD (2009) Dual nature of the adaptive immune system in lampreys. *Nature* 459(7248):796-801.

Guo L, Hu-Li J, Paul WE (2005) Probabilistic regulation in TH2 cells accounts for monoallelic expression of IL-4 and IL-13. *Immunity* 23(1):89-99.

Guo L, Hu-Li J, Zhu J, Watson CJ, Difilippantonio MJ, Pannetier C, Paul WE (2002) In TH2 cells the Il4 gene has a series of accessibility states associated with distinctive probabilities of IL-4 production. *Proc Natl Acad Sci U S A.* 99(16):10623-8.

Gouaillard C, Huchenq-Champagne A, Arnaud J, Chen Cl CL, Rubin B (2001) Evolution of T cell receptor (TCR) alpha beta heterodimer assembly with the CD3 complex. *Eur J Immunol.* 31(12):3798-805.

Granevitze Z, Hillel J, Chen GH, Cuc NT, Feldman M, Eding H, Weigend S. 2007. Genetic diversity within chicken populations from different continents and management histories. *Anim Genet.* 38:576-83.

Granevitze Z, Hillel J, Feldman M, Six A, Eding H, Weigend S (2009) Genetic structure of a wide-spectrum chicken gene pool. *Animal Genetics* Eepub June 3rd.

Griffin DK, Robertson LB, Tempest HG, Skinner BM (2007) The evolution of the avian genome as revealed by comparative molecular cytogenetics. Cytogenet *Genome Res.* 117:64-77.

Gu CC, Yu K, Rao DC (2008) Characterization of LD structures and the utility of HapMap in genetic association studies. *Adv Genet*. 60:407-35.

Gyorfy Z, Ohnemus A, Kaspers B, Duda E, Staeheli P (2003) Truncated chicken interleukin-1beta with increased biologic activity. *J Interferon Cytokine Res*. 23(5):223-8.

Hackett SJ, Kimball RT, Reddy S, Bowie RC, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han KL, Harshman J, Huddleston CJ, Marks BD, Miglia KJ, Moore WS, Sheldon FH, Steadman DW, Witt CC, Yuri T (2008) A phylogenomic study of birds reveals their evolutionary history. *Science* 320(5884):1763-8.

Hansson B, Ljungqvist M, Dawson DA, Mueller JC, Olano-Marin J, Ellegren H, Nilsson JA (2009) Avian genome evolution: insights from a linkage map of the blue tit (*Cyanistes caeruleus*). *Heredity* Eepub August 26th.

Harada A, Azakami H, Kato A (2008) Amyloid fibril formation of hen lysozyme depends on the instability of the C-helix (88-99). B*iosci Biotechnol Biochem*. 72(6):1523-30.

Harris EE, Meyer D (2006) The molecular signature of selection underlying human adaptations. *Am J Phys Anthropol*. 43:89-130.

Hawken RJ, Barris WC, McWilliam SM, Dalrymple BP (2004) An interactive bovine in silico SNP database (IBISS). *Mamm Genome*. 15(10):819-27.

Hayes BJ, Nilsen K, Berg PR, Grindflek E, Lien S (2007) SNP detection exploiting multiple sources of redundancy in large EST collections improves validation rates. *Bioinformatics*. 23(13):1692-3.

He H, Genovese KJ, Nisbet DJ, Kogut MH (2006) Involvement of phosphatidylinositol-phospholipase C in immune response to *Salmonella* lipopolysacharide in chicken macrophage cells (HD11). *Int Immunopharmaco*l. 6:1780-7.

Hebenstreit D, Horejs-Hoeck J, Duschl A. (2005) JAK/STAT-dependent gene regulation by cytokines. *Drug News Perspect*. 18(4):243-9.

Hedges SB (2002) The origin and evolution of model organisms. *Nat Rev Genet*. 3(11):838-49.

Heidari M, Zhang HM, Sharif S (2008) Marek's disease virus induces Th-2 activity during cytolytic infection. *Viral Immunol*. 21:203-14.

Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169(4):2335-52.

Hermisson J (2009) Who believes in whole-genome scans for selection? *Heredity* epub Aug 5th.

Hershey GK, Friedrich MF, Esswein LA, Thomas ML, Chatila TA (1997) The association of atopy with a gain-of-function mutation in the alpha subunit of the interleukin-4 receptor. *N Engl J Med.* 11;337(24):1720-5.

Hiendleder S, Mainz K, Plante Y, Lewalski H (1998) Analysis of mitochondrial DNA indicates that domestic sheep are derived from two different ancestral maternal sources: no evidence for contributions from urial and argali sheep. *J Hered.* 89(2):113-20.

Higgs R, Cormican P, Cahalane S, Allan B, Lloyd AT, Meade K, James T, Lynn DJ, Babiuk LA, O'farrelly C (2006) Induction of a novel chicken Toll-like receptor following Salmonella enterica serovar *Typhimurium* infection. *Infect Immun.* 74:1692-8.

Higuchi M, Matsuo A, Shingai M, Shida K, Ishii A, Funami K, Suzuki Y, Oshiumi H, Matsumoto M, Seya T (2008) Combinational recognition of bacterial lipoproteins and peptidoglycan by chicken Toll-like receptor 2 subfamily. *Dev Comp Immunol.* 32:147-55.

Hillel J, Groenen MA, Tixier-Boichard M, Korol AB, David L, Kirzhner VM, Burke T, Barre-Dirie A, Crooijmans RP, Elo K, Feldman MW, Freidlin PJ, Mäki-Tanila A, Oortwijn M, Thomson P, Vignal A, Wimmers K, Weigend S (2003) Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. *Genet Sel Evol.* 35(5):533-57.

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genetical Research* 8:269-294.

Hoffjan S, Nicolae D, Ober C (2003) Association studies for asthma and atopic diseases: a comprehensive review of the literature. *Respir Res.* 4:14.

Hofmann K, Stoffel W (1993) TMbase - A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler* 374:166.

Holler E, Rupley J, Hess G (1975a) Productive and Unproductive Lysozyme-Chitosaccaride Complexes. Equilibrium Measurements. *Biochemistry* 14:1088-94.

Holler E, Rupley J, Hess G (1975b) Productive and Unproductive Lysozyme-Chitosaccaride Complexes. Kinetic Investigations. *Biochemistry* 14:2377-85.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat Rev Genet.* 10(9):639-50.

Horsnell WG, Cutler AJ, Hoving JC, Mearns H, Myburgh E, Arendse B, Finkelman FD, Owens GK, Erle D, Brombacher F (2007) Delayed goblet cell hyperplasia, acetylcholine receptor expression, and worm expulsion in SMC-specific IL-4Ralpha-deficient mice. *PLoS Pathog.* 3(1):e1.

Hosomichi K, Miller MM, Goto RM, Wang Y, Suzuki S, Kulski JK, Nishibori M, Inoko H, Hanzawa K, Shiina T (2008) Contribution of mutation, recombination, and gene conversion to chicken MHC-B haplotype diversity. *J Immunol.* 181(5):3393-9.

Hou ZC, Xu GY, Su Z, Yang N (2007) Purifying selection and positive selection on the myxovirus resistance gene in mammals and chickens. *Gene* 396(1):188-95.

Howard TD, Meyers DA, Bleecker ER (2000) Mapping susceptibility genes for asthma and allergy. *J Allergy Clin Immunol.* 105(2 Pt 2):S477-81.

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res.* 9(9):868-77.

Hudson RR (1987) Estimating the recombination parameter of a finite population model without selection. *Genet Res.* 50(3):245-50.

Hudson RR (2001) Two-locus sampling distributions and their application. *Genetics* 159(4):1805-17.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337-8.

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111(1):147-64.

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153-159.

Huerta-Sanchez E, Durrett R, Bustamante CD (2008) Population genetics of polymorphism and divergence under fluctuating selection. *Genetics* 178(1):325-37.

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* 99(4):364-73.

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci.* 256(1346):119-24.

Hughes AL, Friedman R (2008) Genome size reduction in the chicken has involved massive loss of ancestral protein-coding genes. *Mol Biol Evol.* 25(12):2681-8.

Hughes AL, Hughes MK (1995) Small genomes for better flyers. *Nature* 377:391.

Hughes AL, Packer B, Welch R, Chanock SJ, Yeager M (2005) High level of functional polymorphism indicates a unique role of natural selection at human immune system loci. *Immunogenetics* 57(11):821-7.

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet*. 32:415-35.

Hurst LD (2009) Fundamental concepts in genetics: genetics and the understanding of selection. *Nat Rev Genet*. 10:83-93.

Ibrahim HR, Thomas U, Pellegrini A (2001) A helix-loop-helix peptide at the upper lip of the active site cleft of lysozyme confers potent antimicrobial activity with membrane permeabilization action. *J Biol Chem*. 276(47):43767-74.

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *Proc Natl Acad Sci U S A*. 101(29):10667-72.

Innan H, Kim Y. 2008. Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics* 179(3):1713-20.

Innan H (2006) Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. *Genetics* 173(3):1725-33.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695-716.

International Chicken Polymorphism Map Consortium (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* 432:717-22.

Iqbal M, Philbin VJ, Withanage GS, Wigley P, Beal RK, Goodchild MJ, Barrow P, McConnell I, Maskell DJ, Young J, Bumstead N, Boyd Y, Smith AL (2005) Identification and functional characterization of chicken toll-like receptor 5 reveals a fundamental role in the biology of infection with Salmonella enterica serovar typhimurium. *Infect Immun*. 73:2344-50.

Itoh Y, Arnold AP (2005) Chromosomal polymorphism and comparative painting analysis in the zebra finch. *Chromosome Res*. 13(1):47-56.

Jackson IJ (1997) Homologous pigmentation mutations in human, mouse and other model organisms. *Hum Mol Genet*. 6(10):1613-24.

Janardhana V, Ford ME, Bruce MP, Broadway MM, O'Neil TE, Karpala AJ, Asif M, Browning GF, Tivendale KA, Noormohammadi AH, Lowenthal JW, Bean AG (2007) IFN-gamma enhances immune responses to *E. coli* infection in the chicken. *J Interferon Cytokine Res*. 27(11):937-46.

Janeway CA Jr, Medzhitov R (2002) Innate immune recognition. *Annu Rev Immunol*. 20:197-216.

Jeffery KJ, Bangham CR (2000) Do infectious diseases drive MHC diversity? *Microbes Infect*. 2(11):1335-41.

Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. *Trends Genet*. 23(11):568-77.

Johnson LN, Phillips DC (1965). Structure of some crystalline lysozyme-inhibitor complexes determined by X-ray analysis at 6 Angstrom resolution. *Nature* 206(986):761-3.

Johnson PL, Slatkin M (2005) Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res*. 16(10):1320-7.

Jung A, Sippel AE, Grez M, Schütz G (1980) Exons encode functional and structural units of chicken lysozyme. *Proc Natl Acad Sci U S A*. 77(10):5759-63.

Junttila IS, Mizukami K, Dickensheets H, Meier-Schellersheim M, Yamane H, Donnelly RP, Paul WE (2008) Tuning sensitivity to IL-4 and IL-13: differential expression of IL-4Ralpha, IL-13Ralpha1, and gammac regulates relative cytokine sensitivity. *J Exp Med.* 205(11):2595-608.

Kaiser P (2007) The avian immune genome – a glass half-full or half-empty? *Cytogenet Genome Res*. 117:221-30.

Kaiser P, Howell MM, Fife M, Sadeyen JR, Salmon N, Rothwell L, Young J, Poh TY, Stevens M, Smith J, Burt D, Swaggerty C, Kogut M (2009) Towards the selection of chickens resistant to *Salmonella* and *Campylobacter* infections. *Bull Mem Acad R Med Belg*. 164(1-2):17-25; discussion 25-6.

Kaiser P, Howell J, Fife M, Sadeyen JR, Salmon N, Rothwell L, Young J, van Diemen P, Stevens M, Poh TY, Jones M, Barrow P, Swaggerty C, Kogut M, Smith J, Burt D (2008) Integrated immunogenomics in the chicken: deciphering the immune response to identify disease resistance genes. *Dev Biol (Basel).* 132:57-66.

Kaiser P, Poh TY, Rothwell L, Avery S, Balu S, Pathania US, Hughes S, Goodchild M, Morrell S, Watson M, Bumstead N, Kaufman J, Young JR (2005) A genomic analysis of chicken cytokines and chemokines. *J Interferon Cytokine Res*. 25:467-84.

Kaiser P, Rothwell L, Goodchild M, Bumstead N (2004) The chicken proinflammatory cytokines interleukin-1beta and interleukin-6: differences in gene structure and genetic location compared with their mammalian orthologues. *Anim Genet*. 35(3):169-75.

Kaiser VB, van Tuinen M, Ellegren H (2007) Insertion events of CR1 retrotransposable elements elucidate the phylogenetic branching order in galliform birds. *Mol Biol Evol*. 24(1):338-47.

Kanginakudru S, Metta M, Jakati RD, Nagaraju J (2008) Genetic evidence from Indian red jungle fowl corroborates multiple domestication of modern day chicken. *BMC Evol Biol*. 8:174.

Kaneko T, Hosokawa H, Yamashita M, Wang CR, Hasegawa A, Kimura MY, Kitajiama M, Kimura F, Miyazaki M, Nakayama T (2007) Chromatin remodeling at the Th2 cytokine gene loci in human type 2 helper T cells. Mol Immunol. 44(9):2249-56.

Karpala AJ, Lowenthal JW, Bean AG (2008) Activation of the TLR3 pathway regulates IFNbeta production in chickens. *Dev Comp Immunol*. 32:435-44.

Kawamoto T, Araki K, Sonoda E, Yamashita YM, Harada K, Kikuchi K, Masutani C, Hanaoka F, Nozaki K, Hashimoto N, Takeda S (2005) Dual roles for DNA polymerase eta in homologous DNA recombination and translesion DNA synthesis. *Mol Cell*. 20(5):793-9.

Kawamura S, Chijiiwa Y, Minematsu T, Fukamizo T, Vårum KM, Torikata T (2008) The role of Arg114 at subsites E and F in reactions catalyzed by hen egg-white lysozyme. *Biosci Biotechnol Biochem*. 72(3):823-32.

Keestra AM, de Zoete MR, van Aubel RA, van Putten JP (2007) The central leucine-rich repeat region of chicken TLR16 dictates unique ligand specificity and species-specific interaction with TLR2. *J Immunol*. 178:7110-9.

Keestra AM, de Zoete MR, van Aubel RA, van Putten JP (2008) Functional characterization of chicken TLR5 reveals species-specific recognition of flagellin. *Mol Immunol*. 45:1298-307.

Keestra AM, van Putten JP (2008) Unique properties of the chicken TLR4/MD-2 complex: selective lipopolysaccharide activation of the MyD88-dependent pathway. *J Immunol*. 181:4354-62.

Kelly JK (1997) A test of neutrality based on interlocus associations. *Genetics* 146(3):1197-206.

Kerje S, Lind J, Schütz K, Jensen P, Andersson L (2003) Melanocortin 1-receptor (MC1R) mutations are associated with plumage colour in chicken. *Anim Genet*. 34:241-8.

Kerje S, Sharma P, Gunnarsson U, Kim H, Bagchi S, Fredriksson R, Schütz K, Jensen P, von Heijne G, Okimoto R, Andersson L (2004) The Dominant white, Dun and Smoky color variants in chicken are associated with insertion/deletion polymorphisms in the PMEL17 gene. *Genetics* 168(3):1507-18.

Kim DK, Hong YH, Park DW, Lamont SJ, Lillehoj HS (2008a) Differential immune-related gene expression in two genetically disparate chicken lines during infection by *Eimeria maxima*. *Dev Biol (Basel)*. 132:131-40.

Kim DK, Lillehoj HS, Hong YH, Park DW, Lamont SJ, Han JY, Lillehoj EP (2008b) Immune-related gene expression in two B-complex disparate genetically inbred Fayoumi chicken lines following *Eimeria maxima* infection. *Poult Sci.* 87(3):433-43.

Kim H, Schmidt CJ, Decker KS, Emara MG (2003) A double-screening method to identify reliable candidate non-synonymous SNPs from chicken EST data. *Anim Genet.* 34(4):249-54.

Kim PM, Korbel JO, Gerstein MB (2007) Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A.* 104(51):20274-9.

Kim Y (2006) Allele frequency distribution under recurrent selective sweeps. *Genetics* 172(3):1967-78.

Kimura M (1979) Model of effectively neutral mutations in which selective constraint is incorporated. *Proc Natl Acad Sci U S A.* 76(7):3440-3444.

Kimura M, Crow JF (1964). The number of alleles that can be maintained in a finite population. *Genetics* 49:725-38.

Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc Natl Acad Sci U S A.* 75(6):2868-72.

Klein-Seetharaman J, Oikawa M, Grimshaw SB, Wirmer J, Duchardt E, Ueda T, Imoto T, Smith LJ, Dobson CM, Schwalbe H (2002) Long-range interactions within a nonnative protein. *Science* 295:1719-22.

Ko JH, Jin HK, Asano A, Takada A, Ninomiya A, Kida H, Hokiyama H, Ohara M, Tsuzuki M, Nishibori M, Mizutani M, Watanabe T (2002) Polymorphisms and the differential antiviral activity of the chicken Mx gene. *Genome Res.* 12(4):595-601.

Kogut MH, Iqbal M, He H, Philbin V, Kaiser P, Smith A (2005a) Expression and function of Toll-like receptors in chicken heterophils. *Dev Comp Immunol.* 29:791-807.

Kogut MH, He H, Kaiser P (2005b) Lipopolysaccharide binding protein/CD14/ TLR4-dependent recognition of S*almonella* LPS induces the functional activation of chicken heterophils and up-regulation of pro-inflammatory cytokine and chemokine gene expression in these cells. *Anim Biotechnol.* 16(2):165-81.

Kogut MH, Rothwell L, Kaiser P (2005c) IFN-gamma priming of chicken heterophils upregulates the expression of proinflammatory and Th1 cytokine mRNA following receptor-mediated phagocytosis of *Salmonella enterica* serovar enteritidis. *J Interferon Cytokine Res.* 25(2):73-81.

Kogut MH, Swaggerty C, He H, Pevzner I, Kaiser P (2006) Toll-like receptor agonists stimulate differential functional activation and cytokine and chemokine gene expression in heterophils isolated from chickens with differential innate responses. *Microbes Infect.* 8(7):1866-74.

Kolm N, Stein RW, Mooers AØ, Verspoor JJ, Cunningham EJ (2007) Can sexual selection drive female life histories? A comparative study on Galliform birds. J Evol Biol. 20(2):627-38.

Koskela K, Kohonen P, Salminen H, Uchida T, Buerstedde JM, Lassila O (2004) Identification of a novel cytokine-like transcript differentially expressed in avian gammadelta T cells. *Immunogenetics* 55(12):845-54.

Kozma N, Halasz M, Polgar B, Poehlmann TG, Markert UR, Palkovics T, Keszei M, Par G, Kiss K, Szeberenyi J, Grama L, Szekeres-Bartho J (2006) Progesterone-induced blocking factor activates STAT6 via binding to a novel IL-4 receptor. *J Immunol.* 176(2):819-26.

Kreitman M, Di Rienzo A (2004) Balancing claims for balancing selection. *Trends Genet.* 20:300-4.

Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, Brosius J, Schmitz J (2007) Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). *BMC Evol Biol.* 7:190.

Kruse S, Japha T, Tedner M, Sparholt SH, Forster J, Kuehr J, Deichmann KA (1999) The polymorphisms S503P and Q576R in the interleukin-4 receptor alpha gene are associated with atopy and influence the signal transduction. *Immunology* 96(3):365-71.

Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. *Nature* 392(6679):917-20.

Landi S, Bottari F, Gemignani F, Gioia-Patricola L, Guino E, Osorio A, de Oca J, Capella G, Canzian F, Moreno V; Bellvitge Colorectal Cancer Study Group (2007) Interleukin-4 and interleukin-4 receptor polymorphisms and colorectal cancer risk. *Eur J Cancer* 43(4):762-8.

Laun K, Coggill P, Palmer S, Sims S, Ning Z, Ragoussis J, Volpi E, Wilson N, Beck S, Ziegler A, Volz A (2006). The leukocyte receptor complex in chicken is characterized by massive expansion and diversification of immunoglobulin-like Loci. *PLoS Genet.* 2(5):e73.

Lee B, Hong T, Byun SJ, Woo T, Choi YJ (2007) ESTpass: a web-based server for processing and annotating expressed sequence tag (EST) sequences. *Nucleic Acids Res.* 35(Web Server issue):W159-62.

Lee BT, Tan TW, Ranganathan S (2003) MGAlignIt: A web service for the alignment of mRNA/EST and genomic sequences. *Nucleic Acids Res.* 31(13):3533-6.

Lee MA, Keane OM, Glass BC, Manley TR, Cullen NG, Dodds KG, McCulloch AF, Morris CA, Schreiber M, Warren J, Zadissa A, Wilson T, McEwan JC (2006) Establishment of a pipeline to analyse non-synonymous SNPs in *Bos taurus*. *BMC Genomics*. 7:298.

Lendahl U, Zimmerman LB, McKay RD (1990) CNS stem cells express a new class of intermediate filament protein. *Cell*. 60(4):585-95.

Leulier F, Lemaitre B (2008) Toll-like receptors--taking an evolutionary approach. *Nat Rev Genet*. 9:165-78.

Leveque G, Forgetta V, Morroll S, Smith AL, Bumstead N, Barrow P, Loredo-Osti JC, Morgan K, Malo D (2003) Allelic variation in TLR4 is linked to susceptibility to *Salmonella enterica* serovar Typhimurium infection in chickens. *Infect Immun*. 71:1116-24.

Li WH (1976) A Mixed Model of Mutation for Electrophoretic Identity of Proteins within and between Populations. *Genetics* 83(2):423-432.

Li XY, Qu LJ, Yao JF, Yang N (2006) Skewed allele frequencies of an Mx gene mutation with potential resistance to avian influenza virus in different chicken populations. *Poult Sci*. 85(7):1327-9.

Li YD, Xie ZY, Du YL, Zhou Z, Mao XM, Lv LX, Li YQ (2009) The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene* 436:8-11.

Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451-1452.

Ling MK, Lagerström MC, Fredriksson R, Okimoto R, Mundy NI, Takeuchi S, Schiöth HB (2003) Association of feather colour with constitutively active melanocortin 1 receptors in chicken. *Eur J Biochem*. 270(7):1441-9.

Link H (1998) The cytokine storm in multiple sclerosis. *Mult Scler*. 4(1):12-5.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJ, van Oudenaarden A, Barton DB, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ (2009) Population genomics of domestic and wild yeasts. *Nature* 458:337-41.

Liu B, Liao J, Rao X, Kushner SA, Chung CD, Chang DD, Shuai K (1998) Inhibition of Stat1-mediated gene activation by PIAS1. *Proc Natl Acad Sci U S A*., 95(18):10626-31.

Liu X, Beaty TH, Deindl P, Huang SK, Lau S, Sommerfeld C, Fallin MD, Kao WH, Wahn U, Nickel R (2004) Associations between specific serum IgE response and 6 variants within the genes IL4, IL13, and IL4RA in German children: the German Multicenter Atopy Study. *J Allergy Clin Immunol.* 113(3):489-95.

Liu YP, Wu GS, Yao YG, Miao YW, Luikart G, Baig M, Beja-Pereira A, Ding ZL, Palanichamy MG, Zhang YP (2006) Multiple maternal origins of chickens: out of the Asian jungles. *Mol Phylogenet Evol.* 38:12–19.

Lochmiller RL, Deerenberg C (2000) Trade-offs in evolutionary immunology: just what is the cost of immunity? *Oikos* 88(1):87-98.

Long JE, Huang LN, Qin ZQ, Wang WY, Qu D (2004) IFN-gamma increases efficiency of DNA vaccine in protecting ducks against infection. *World J Gastroenterol.* 11(32):4967-73.

Lozano F, Places L, Vilà JM, Padilla O, Arman M, Gimferrer I, Suárez B, López de la Iglesia A, Miserachs N, Vives J (2001) Identification of a novel single-nucleotide polymorphism (Val554Ile) and definition of eight common alleles for human IL4RA exon 11. *Tissue Antigens* 57(3):216-20.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151-5.

Lynn DJ, Lloyd AT, O'Farrelly C (2003) In silico identification of components of the Toll-like receptor (TLR) signaling pathway in clustered chicken expressed sequence tags (ESTs). *Vet Immunol Immunopathol.* 93(3-4):177-84.

Lyons EJ, Amos W, Berkley JA, Mwangi I, Shafi M, Williams TN, Newton CR, Peshu N, Marsh K, Scott JA, Hill AV (2009a) Homozygosity and risk of childhood death due to invasive bacterial disease. *BMC Med Genet.* 10(1):55.

Lyons EJ, Frodsham AJ, Zhang L, Hill AV, Amos W (2009b) Consanguinity and susceptibility to infectious diseases in humans. *Biol Lett.* epub Jun 11th.

Madge LA, Pober JS (2000) A phosphatidylinositol 3-kinase/Akt pathway, activated by tumor necrosis factor or interleukin-1, inhibits apoptosis but does not activate NFkappaB in human endothelial cells. *J Biol Chem.* 275(20):15458-65.

Madge S, McGowan P, Kirwan GM (2002) "Pheasants, Partridges and Grouse: a guide to pheasants, partridges, quails, grouse, guineafowl, buttonquails and sandgrouse of the world". 1st edition. Christopher Helm Press, London.

Malek M, Hasenstein JR, Lamont SJ (2004) Analysis of chicken TLR4, CD28, MIF, MD-2, and LITAF genes in a *Salmonella* enteritidis resource population. *Poult Sci.* 83:544-9.

Marshall Graves JA (2009) Sex determination: Birds do it with a Z gene. *Nature* 461:177-178.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351(6328):652-4.

McKenzie AN, Culpepper JA, de Waal Malefyt R, Brière F, Punnonen J, Aversa G, Sato A, Dang W, Cocks BG, Menon S, De Vries SE, Banchereau J, Zurawski G (1993) Interleukin 13, a T-cell-derived cytokine that regulates human monocyte and B-cell function. *Proc Natl Acad Sci U S A.* 90(8):3735-9.

McPherson JD, Dodgson J, Krumlauf R, Pourquiè O. Proposal to sequence the genome of the chicken. Trans-NIH Gallus Initiative. www.nih.gov/science/models/gallus/ChickenGenomeWhitePaper.pdf.

McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160(3):1231-41.

Medzhitov R, Janeway C Jr (2000) Innate immunity. *N Engl J Med.* 343(5):338-44.

Mege JL, Meghari S, Honstettre A, Capo C, Raoult D (2006) The two faces of interleukin 10 in human infectious diseases. *Lancet Infect Dis.* 6(9):557-69.

Merilä J (2009) Genetic Constraints on Adaptation? *Science* 325:1212-3.

Michaux C, Pouyez J, Wouters J, Privé GG (2008) Protecting role of cosolvents in protein denaturation by SDS: a structural study. *BMC Structural Biology* 8:29.

Minvielle F, Cecchi T, Passamonti P, Gourichon D, Renieri C (2009) Plumage colour mutations and melanins in the feathers of the Japanese quail: a first comparison. *Anim Genet.* Epub Jun 3rd.

Moeller DA, Tiffin P (2008) Geographic variation in adaptation at the molecular level: a case study of plant immunity genes. *Evolution* 62(12):3069-81.

Mohanty TR, Seo KS, Park KM, Choi TJ, Choe HS, Baik DH, Hwang IH (2008) Molecular variation in pigmentation genes contributing to coat colour in native Korean Hanwoo cattle. *Anim Genet.* 39(5):550-3.

Monte M, Benetti R, Buscemi G, Sandy P, Del Sal G, Schneider C (2003) The cell cycle-regulated protein human GTSE-1 controls DNA damage-induced apoptosis by affecting p53 function. *J Biol Chem.* 278(32):30356-64.

Moult J, Yonath A, Traub W, Smilansky A, Podjarny A, Rabinovich D, Saya A (1976) The structure of triclinic lysozyme at 2-5 A resolution. *J. Mol. Biol.* 100(2): 179-195.

Msoffe PL, Minga UM, Mtambo MM, Gwakisa PS, Olsen JE (2006) Differences in resistance to Salmonella enterica serovar Gallinarum infection among indigenous local chicken ecotypes in Tanzania. *Avian Pathol.* 35(4):270-6.

Muchadeyi FC, Eding H, Simianer H, Wollny CB, Groeneveld E, Weigend S (2008) Mitochondrial DNA D-loop sequences suggest a Southeast Asian and Indian origin of Zimbabwean village chickens. *Anim Genet.* 39(6):615-22.

Muir WM, Wong GK, Zhang Y, Wang J, Groenen MA, Crooijmans RP, Megens HJ, Zhang H, Okimoto R, Vereijken A, Jungerius A, Albers GA, Lawley CT, Delany ME, MacEachern S, Cheng HH (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proc Natl Acad Sci U S A.* 105(45):17312-7.

Mulard H, Danchin E, Talbot SL, Ramey AM, Hatch SA, White JF, Helfenstein F, Wagner RH (2009) Evidence that pairing with genetically similar mates is maladaptive in a monogamous bird. *BMC Evol Biol.* 9:147.

Mundy NI (2005) A window on the genetics of evolution: MC1R and plumage colouration in birds. *Proc Biol Sci.* 272(1573):1633-40.

Mwacharo JM, Nomura K, Hanada H, Jianlin H, Hanotte O, Amano T (2007) Genetic relationships among Kenyan and other East African indigenous chickens. *Anim Genet.* 38(5):485-90.

Nadeau NJ, Minvielle F, Mundy NI (2006)Association of a Glu92Lys substitution in MC1R with extended brown in Japanese quail (*Coturnix japonica*). *Anim Genet.* 37(3):287-9.

Nakayama T, Yamashita M, Kimura M, Hasegawa A, Omori M, Inami M, Motohashi S, Kitajima M, Hashimoto K, Hosokawa H, Shinnakasua R (2005) Chromatin remodeling of the Th2 cytokine gene loci. International Congress Series 1285:137-144.

Nerren JR, Swaggerty CL, MacKinnon KM, Genovese KJ, He H, Pevzner I, Kogut MH (2009) Differential mRNA expression of the avian-specific toll-like receptor 15 between heterophils from Salmonella-susceptible and -resistant chickens. *Immunogenetics* 61:71-7.

Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31(13):3812-4.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* Epub Aug 16th.

Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*. 25(14):2745-51.

Nie W, O'Brien PC, Ng BL, Fu B, Volobouev V, Carter NP, Ferguson-Smith MA, Yang F (2009) Avian comparative genomics: reciprocal chromosome painting between domestic chicken (*Gallus gallus*) and the stone curlew (*Burhinus oedicnemus*, Charadriiformes)--an atypical species with low diploid number. *Chromosome Res*. 17:99-113.

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929-936.

Nishibori M, Shimogiri T, Hayashi T, Yasue H (2005) Molecular evidence for hybridization of species in the genus *Gallus* except for *Gallus varius*. *Animal Genetics* 36:367-75.

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 302(1):205-17.

Nozawa M, Suzuki Y, Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A*. 106(16):6700-5.

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res*. 22:201-204.

Oka T, Ino Y, Nomura K, Kawashima S, Kuwayama T, Hanada H, Amano T, Takada M, Takahata N, Hayashi Y, Akishinonomiya F (2007) Analysis of mtDNA sequences shows Japanese native chickens have multiple origins. *Animal Genetics* 38:287-93.

Okamura M, Lillehoj HS, Raybourne RB, Babu US, Heckert RA (2004) Cell-mediated immune responses to a killed Salmonella enteritidis vaccine: lymphocyte proliferation, T-cell changes and interleukin-6 (IL-6), IL-1, IL-2, and IFN-gamma production. *Comp Immunol Microbiol Infect Dis*. 27(4):255-72.

O'Neill AM, Livant EJ, Ewald SJ (2009) The chicken BF1 (classical MHC class I) gene shows evidence of selection for diversity in expression and in promoter and signal peptide regions. *Immunogenetics* 61(4):289-302.

Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV (2007) Origin of avian genome size and structure in non-avian dinosaurs. *Nature* 446:180-4.

Ortutay C, Vihinen M (2006) Immunome: a reference set of genes and proteins for systems biology of the human immune system. *Cell Immunol*. 244:87-9.

Ots I, Kerimov AB, Ivankina EV, Ilyina TA, Hõrak P (2001) Immune challenge affects basal metabolic activity in wintering great tits. *Proc Biol Sci.* 268(1472):1175-81.

Ozoe A, Isobe N, Yoshimura Y (2009) Expression of Toll-like receptors (TLRs) and TLR4 response to lipopolysaccharide in hen oviduct. *Vet Immunol Immunopathol.* 127:259-68.

Pagani F, Raponi M, Baralle FE (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci U S A.* 102(18):6368-72.

Pang JF, Kluetsch C, Zou XJ, Zhang A, Luo LY, Angleby H, Ardalan A, Ekström C, Sköllermo A, Lundeberg J, Matsumura S, Leitner T, Zhang JP, Savolainen P (2009) mtDNA Data Indicates a Single Origin for Dogs South of Yangtze River, less than 16,300 Years Ago, from Numerous Wolves. *Molecular Biology & Evolution* Epub Sept 1st.

Park D, Hochreiter-Hufford A, Ravichandran KS (2009) The phosphatidylserine receptor TIM-4 does not mediate direct signaling. *Curr Biol.* 19(4):346-51.

Pellegrini A, Thomas U, Bramaz N, Klauser S, Hunziker P, von Fellenberg R (1997) Identification and isolation of a bactericidal domain in chicken egg white lysozyme. *J Appl Microbiol.* 82(3):372-8.

Pennings PS, Hermisson J (2006b) Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. *Mol Biol Evol.* 23(5):1076-84.

Pennings PS, Hermisson J (2006c) Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.* 2(12):e186.

Pepys MB, Hawkins PN, Booth DR, Vigushin DM, Tennent GA, Soutar AK, Totty N, Nguyen O, Blake CC, Terry CJ, Feest TG, Zalin AM, Hsuan JJ (1993) Human lysozyme gene mutations cause hereditary systemic amyloidosis. *Nature* 362(6420):553-7.

Pereira SL, Baker AJ, Wajntal A (2002) Combined Nuclear and Mitochondrial DNA Sequences Resolve Generic Relationships within the Cracidae (Galliformes, Aves). *Systematic Biology* 51:946-958.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19(5):651-2.

Pointer MA, Mundy NI (2008) Testing whether macroevolution follows microevolution: are colour differences among swans (*Cygnus*) attributable to variation at the MCIR locus? *BMC Evol Biol.* 8:249.

Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge D, Rotter JI, Torres E, Taylor KD, Goldstein DB, Reich D (2008) Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet.* 83(1):132-5.

Prieur DJ, Olson HM, Young DM (1974) Lysozyme deficiency-an inherited disorder of rabbits. *Am J Pathol.* 77(2):283-98.

Pryke SR, Andersson S, Lawes MJ, Pipera SE (2002) Carotenoid status signaling in captive and wild red-collared widowbirds: independent effects of badge size and color. *Behavioral Ecology* 13:622-631.

Pryke SR (2009) Is red an innate or learned signal of aggression and intimidation? *Animal Behaviour* Epub 27th June.

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. *Evolution* 59(11):2312-23.

Puri RK, Aggarwal BB (1992) Human immunodeficiency virus type 1 tat gene up-regulates interleukin 4 receptors on a human B-lymphoblastoid cell line. Cancer Res. 52(13):3787-90.

Quesada H, Ramirez UE, Rozas J, Aguade M (2006) Large-scale adaptive hitchhiking upon high recombination in *Drosophila simulans*. *Genetics* 165(2):895-900.

Råberg L, Vestberg M, Hasselquist D, Holmdahl R, Svensson E, Nilsson JA (2002) Basal metabolic rate and the evolution of the adaptive immune system. *Proc Biol Sci.* 269(1493):817-21.

Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30(17):3894-900.

Rauw F, Lambrecht B, van den Berg T (2007) Pivotal role of ChIFNgamma in the pathogenesis and immunosuppression of infectious bursal disease. *Avian Pathol.* 36(5):367-74.

Razafindraibe H, Mobegi VA, Ommeh SC, Rakotondravao ML, Bjørnstad G, Hanotte O, Jianlin H (2008) Mitochondrial DNA origin of indigenous malagasy chicken. *Ann N Y Acad Sci.* 1149:77-9.

Reed FA, Tishkoff SA (2006) Positive selection can create false hotspots of recombination. *Genetics* 172(3):2011-4.

Reid JM, Arcese P, Keller LF, Elliott KH, Sampson L, Hasselquist D (2007) Inbreeding effects on immune response in free-living song sparrows (*Melospiza melodia*). *Proc Biol Sci.* 274(1610):697-706.

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol.* 9(3):552-69.

Ronald J, Akey JM (2005) Genome-wide scans for loci under selection in humans. *Hum Genomics*. 2(2):113-25.

Rosenblum EB, Hoekstra HE, Nachman MW (2004) Adaptive reptile color variation and the evolution of the Mc1r gene. *Evolution* 58(8):1794-808.

Roulin A, Jungi TW, Pfister H, Dijkstra C (2000) Female barn owls (*Tyto alba*) advertise good genes. *Proc Biol Sci*. 267(1446):937-41.

Roulin A, Riols C, Dijkstra C, Ducrest AL (2001) Female plumage spottiness and parasite resistance in the barn owl (*Tyto alba*). Behavioral Ecology 12:103-110.

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15(2):174-5.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19(18):2496-7.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. *Science* 312(5780):1614-20.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG (2007) Dynamic evolution of the innate immune system in Drosophila. *Nat Genet.* 39:1461-8.

Sadeyen JR, Trotereau J, Velge P, Marly J, Beaumont C, Barrow PA, Bumstead N, Lalmanach AC (2004) Salmonella carrier state in chicken: comparison of expression of immune response genes between susceptible and resistant animals. *Microbes Infect*. 6(14):1278-86.

Saks L, Ots I, Hõrak P (2003) Carotenoid-based plumage coloration of male greenfinches reflects health and immunocompetence. *Oecologia.* 134(3):301-7.

Salamini F, Ozkan H, Brandolini A, Schäfer-Pregl R, Martin W (2002) Genetics and geography of wild cereal domestication in the near east. *Nat Rev Genet.* 3(6):429-41.

Sawyer SL, Wu LI, Emerman M, Malik HS (2005) Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proc Natl Acad Sci U S A.* 102(8):2832-7.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15(11):1576-83.

Schiöth HB, Raudsepp T, Ringholm A, Fredriksson R, Takeuchi S, Larhammar D, Chowdhary BP (2003) Remarkable synteny conservation of melanocortin receptors in chicken, human, and other vertebrates. *Genomics* 81(5):504-9.

Schlenke TA, Begun DJ (2003) Natural selection drives Drosophila immune system evolution. *Genetics* 164(4):1471-80.

Schneider S., Roessli D., Excoffier L. (2000) Arlequin: a software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

Schoenborn JR, Wilson CB (2007) Regulation of interferon-gamma during innate and adaptive immune responses. *Adv Immunol.* 96:41-101.

Sen GC (2001) Viruses and interferons. *Annu Rev Microbiol.*55:255-81.

Seyama T, Ko JH, Ohe M, Sasaoka N, Okada A, Gomi H, Yoneda A, Ueda J, Nishibori M, Okamoto S, Maeda Y, Watanabe T (2006) Population research of genetic polymorphism at amino acid position 631 in chicken Mx protein with differential antiviral activity. *Biochem Genet.* 44(9-10):437-48.

Shaughnessy RG, Meade KG, Cahalane S, Allan B, Reiman C, Callanan JJ, O'Farrelly C (2009) Innate immune gene expression differentiates the early avian intestinal response between *Salmonella* and *Campylobacter*. *Vet Immunol Immunopathol.* epub Jun 21st.

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol.* 5(6):704-16.

Shiina T, Shimizu S, Hosomichi K, Kohara S, Watanabe S, Hanzawa K, Beck S, Kulski JK, Inoko H (2004) Comparative genomic analysis of two avian (quail and chicken) MHC regions. *J Immunol.* 172(11):6751-63.

Shirakawa I, Deichmann KA, Izuhara I, Mao I, Adra CN, Hopkin JM (2000) Atopy and asthma: genetic variants of IL-4 and IL-13 signalling. *Immunol Today* 21(2):60-4.

Shuai K, Liu B (2005) Regulation of gene-activation pathways by PIAS proteins in the immune system. *Nat Rev Immunol.* 5(8):593-605.

Silva P, Guan X, Ho-Shing O, Jones J, Xu J, Hui D, Notter D, Smith E (2008) Mitochondrial DNA-based analysis of genetic variation and relatedness among Sri Lankan indigenous chickens and the Ceylon junglefowl (*Gallus lafayeti*). *Anim Genet.* Epub Oct 20th.

Simon A, Fäh J, Haller O, Staeheli P (1991) Interferon-regulated Mx genes are not responsive to interleukin-1, tumor necrosis factor, and other cytokines. *J. Virol.* 65(2):968-71.

Sippel AE, Land H, Lindenmaier W, Nguyen-Huu MC, Wurtz T, Timmis KN, Giesecke K, Schütz G (1978) Cloning of chicken lysozyme structural gene sequences synthesized in vitro. *Nucleic Acids Research* 5:3275-94.

Smith CA, Roeszler KN, Ohnesorg T, Cummins DM, Farlie PG, Doran TJ, Sinclair AH (2009) The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature* 461:267-271.

Smith CK, Kaiser P, Rothwell L, Humphrey T, Barrow PA, Jones MA (2004) Campylobacter jejuni-induced cytokine responses in avian cells. *Infect Immun.* (4):2094-100.

Smith NG, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. *Nature* 415(6875):1022-4.

Soldevila M, Calafell F, Helgason A, Stefánsson K, Bertranpetit J (2005) Assessing the signatures of selection in PRNP from polymorphism data: results support Kreitman and Di Rienzo's opinion. *Trends Genet.* 21:389-91.

Spangelo BL, Farrimond DD, Pompilius M, Bowman KL (2000) Interleukin-1 beta and thymic peptide regulation of pituitary and glial cell cytokine expression and cellular proliferation. *Ann N Y Acad Sci.* 917:597-607.

Staeheli P, Puehler F, Schneider K, Göbel TW, Kaspers B (2001) Cytokines of birds: conserved functions--a largely different look. *J Interferon Cytokine Res.* 21(12):993-1010.

Stapley J, Birkhead TR, Burke T, Slate J (2008) A linkage map of the zebra finch *Taeniopygia guttata* provides new insights into avian genome evolution. *Genetics* 179(1):651-67.

Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat Genet.* 38(3):375-81.

Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 68: 978-989.

Stollenwerk MM, Lindholm MW, Pörn-Ares MI, Larsson A, Nilsson J, Ares MP (2005) Very low-density lipoprotein induces interleukin-1beta expression in macrophages. *Biochem Biophys Res Commun.* 335(2):603-8.

Storey AA, Ramírez JM, Quiroz D, Burley DV, Addison DJ, Walter R, Anderson AJ, Hunt TL, Athens JS, Huynen L, Matisoo-Smith EA (2007) Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile. *Proc Natl Acad Sci U S A.* 104(25):10335-9.

Storz JF (2009) Genome evolution: gene duplication and the resolution of adaptive conflict. *Heredity* 102:99-100.

Strynadka NCJ, James MNG (1991) Lysozyme Revisited: Crystallographic Evidence for Distortion of an N-Acetylmuramic Acid Residue Bound in Site D. *Journal of Molecular Biology* 220(2):401-424.

Sugano S, Stoeckle MY, Hanafusa H (1987) Transformation by Rous sarcoma virus induces a novel gene with homology to a mitogenic platelet protein. *Cell* 49:321-8.

Swaggerty CL, Pevzner IY, He H, Genovese KJ, Nisbet DJ, Kaiser P, Kogut MH (2009) Selection of broilers with improved innate immune responsiveness to reduce on-farm infection by foodborne pathogens. *Foodborne Pathog Dis*. 6(7):777-83.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics* 168(3):1457-65.

Szarski H (1976) Cell size and nuclear DNA content in vertebrates. *Int Rev Cytol*. 44:93–209.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105(2):437-60.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585-95.

Tajima F (1993) Statistical analysis of DNA polymorphism. *Jpn. J. Genet*. 68(6):567-95.

Tajima F, Mukai T (1990) Some consideration on diversifying selection. *Jpn J Genet*. 65(4):193-200.

Takeuchi S, Suzuki S, Hirose S, Yabuuchi M, Sato C, Yamamoto H, Takahashi S (1996a) Molecular cloning and sequence analysis of the chick melanocortin 1-receptor gene. *Biochim Biophys Acta*. 1306(2-3):122-6.

Takeuchi S, Suzuki H, Yabuuchi M, Takahashi S (1996b) A possible involvement of melanocortin 1-receptor in regulating feather color pigmentation in the chicken. *Biochim Biophys Acta*. 1308(2):164-8.

Takeuchi S, Teshigawara K, Takahashi S (1999) Molecular cloning and characterization of the chicken pro-opiomelanocortin (POMC) gene. *Biochim Biophys Acta*. 1450(3):452-9.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*. 24(8):1596-9.

Tan CP, McKee KK, Weinberg DH, MacNeil T, Palyha OC, Feighner SD, Hreniuk DL, Van Der Ploeg LH, MacNeil DJ, Howard AD (1999) Molecular analysis of a new splice variant of the human melanocortin-1 receptor. *FEBS Lett.* 451(2):137-41.

Tang J, Leunissen JA, Voorrips RE, van der Linden CG, Vosman B (2008) HaploSNPer: a web-based allele and SNP detection tool. *BMC Genet.* 9:23.

Temperley ND, Berlin S, Paton IR, Griffin DK, Burt DW (2008) Evolution of the chicken Toll-like receptor gene family: a story of gene gain and gene loss. *BMC Genomics* 9:62.

Theron E, Hawkins K, Bermingham E, Ricklefs RE, Mundy NI (2001) The molecular basis of an avian plumage polymorphism in the wild: a melanocortin-1-receptor point mutation is perfectly associated with the melanic plumage morph of the bananaquit, *Coereba flaveola*. *Curr Biol.* 11(8):550-7.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13(9):2129-41.

Tiemann I, Rehkämper G (2009) Effect of artificial selection on female choice among domesticated chickens *Gallus gallus* f.d. *Poult Sci.* 88(9):1948-54.

Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnol J.* 7(4):334-46.

Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT, Cunningham P, Chamberlain AT, Sykes BC, Bradley DG (2001) Genetic evidence for Near-Eastern origins of European cattle. *Nature* 410(6832):1088-91.

Uy JA, Moyle RG, Filardi CE, Cheviron ZA (2009) Difference in Plumage Color Used in Species Recognition between Incipient Species Is Linked to a Single Amino Acid Substitution in the Melanocortin-1 Receptor. *Am Nat.* Epub Jan 1st.

Valverde P, Healy E, Jackson I, Rees JL, Thody AJ (1995) Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat Genet.* 11(3):328-30.

Vinther J, Briggs DE, Clarke J, Mayr G, Prum RO (2009) Structural coloration in a fossil feather. *Biol Lett.* Epub Aug 26th.

Wada K, Howard JT, McConnell P, Whitney O, Lints T, Rivas MV, Horita H, Patterson MA, White SA, Scharff C, Haesler S, Zhao S, Sakaguchi H, Hagiwara M, Shiraki T, Hirozane-Kishikawa T, Skene P, Hayashizaki Y, Carninci P, Jarvis ED (2006) A molecular neuroethological approach for identifying and characterizing a cascade of behaviorally regulated genes. *Proc Natl Acad Sci U S A.* 103(41):15212-7.

Wakeley J (2008) Conditional gene genealogies under strong purifying selection. *Mol Biol Evol.* 25(12):2615-26.

Wang JP, Lindsay BG, Leebens-Mack J, Cui L, Wall K, Miller WC, dePamphilis CW (2004) EST clustering error evaluation and correction. *Bioinformatics* 20(17):2973-84.

Wang X, Rosa AJ, Oliverira HN, Rosa GJ, Guo X, Travnicek M, Girshick T (2006) Transcriptome of local innate and adaptive immunity during early phase of infectious bronchitis viral infection. *Viral Immunol.* 19:768-74.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7:256-276.

Watterson GA (1978) An Analysis of Multi-Allelic Data. *Genetics* 88:171–179.

Webster MT, Axelsson E, Ellegren H (2006) Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol.* 23(6):1203-16.

Weining KC, Sick C, Kaspers B, Staeheli P (1998) A chicken homolog of mammalian interleukin-1 beta: cDNA cloning and purification of active recombinant protein. *Eur J Biochem.* 258(3):994-1000.

West B, Zhou BX (1989) Did chickens go north? New evidence for domestication. *World's Poultry Science Journal.* 45(3):205-218.

Williams EJ, Pal C, Hurst LD (2000) The molecular evolution of signal peptides. *Gene* 253(2):313-22.

Williamson K (2000) Did chickens go west? In "The origins and development of African livestock: archaeology, genetics, linguistics and ethnography". 1st edition. Edited by Blench RM, McDonald KC. Routledge. 368-448.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R (2009) Localizing recent adaptive evolution in the human genome. *PLoS Genet.* 3(6):e90.

Wilson JN, Rockett K, Keating B, Jallow M, Pinder M, Sisay-Joof F, Newport M, Kwiatkowski D (2006) A hallmark of balancing selection is present at the promoter region of interleukin 10. *Genes Immun.* 7(8):680-3.

Wilson K.P., Malcolm B.A., Matthews B.W (1992) Structural and thermodynamic analysis of compensating mutations within the core of chicken egg white lysozyme. *Journal of Biological Chemistry.* 267:10842-9.

Wlasiuk G, Khan S, Switzer WM, Nachman MW (2009) A history of recurrent positive selection at the toll-like receptor 5 in primates. Mol Biol Evol. 26(4):937-49.

Wong JP, Christopher ME, Viswanathan S, Karpoff N, Dai X, Das D, Sun LQ, Wang M, Salazar AM (2009) Activation of toll-like receptor signaling pathway for protection against influenza virus infection. *Vaccine* 27(25-26):3481-3.

Woolley SM, Posada D, Crandall KA (2008) A comparison of phylogenetic network methods using computer simulation. *PLoS One*. 3(4):e1913.

Worley K, Gillingham M, Jensen P, Kennedy LJ, Pizzari T, Kaufman J, Richardson DS (2008) Single locus typing of MHC class I and class II B loci in a population of red jungle fowl. *Immunogenetics* 60(5):233-47.

Wrensch M, Wiencke JK, Wiemels J, Miike R, Patoka J, Moghadassi M, McMillan A, Kelsey KT, Aldape K, Lamborn KR, Parsa AT, Sison JD, Prados MD (2006) Serum IgE, tumor epidermal growth factor receptor expression, and inherited polymorphisms associated with glioma survival. *Cancer Res*. 66(8):4531-41.

Wright, S (1951) The genetical structure of populations. *Annals of Eugenics* 15:323-354.

Wu X, Di Rienzo A, Ober C (2001) A population genetics study of single nucleotide polymorphisms in the interleukin 4 receptor alpha (IL4RA) gene. *Genes Immun*. 2(3):128-34.

Wu ZL, Li XL, Liu YQ, Gong YF, Liu ZZ, Wang XJ, Xin TR, Ji Q (2006) The red head and neck of Boer goats may be controlled by the recessive allele of the MC1R gene. *Anim. Res*. 55:313-322.

Xing Z, Cardona CJ, Li J, Dao N, Tran T, Andrada J (2008) Modulation of the immune responses in chickens by low-pathogenicity avian influenza virus H9N2. *J Gen Virol*. 89:1288-99.

Xing Z, Cardona CJ, Li J, Dao N, Tran T, Andrada J (2008) Modulation of the immune responses in chickens by low-pathogenicity avian influenza virus H9N2. *J Gen Virol*. 89(Pt 5):1288-99.

Yamashita M, Ukai-Tadenuma M, Kimura M, Omori M, Inami M, Taniguchi M, Nakayama T (2002) Identification of a conserved GATA3 response element upstream proximal from the interleukin-13 gene locus. *J Biol Chem*. 277(44):42399-408.

Yamashita M, Ukai-Tadenuma M, Miyamoto T, Sugaya K, Hosokawa H, Hasegawa A, Kimura M, Taniguchi M, DeGregori J, Nakayama T (2004) Essential role of GATA3 for the maintenance of type 2 helper T (Th2) cytokine production and chromatin remodelling at the Th2 cytokine gene loci. *J Biol Chem*. 279(26):26983-90.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*. 13:555-556.

Yang Z (2002) Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev*. 12:688–694.

Yang Z, Lu Z, Wang A (2001) Study of adaptive mutations in *Salmonella* typhimurium by using a super-repressing mutant of a trans regulatory gene purR. *Mutat Res*. 484(1-2):95-102.

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19(6):908-17.

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22(4):1107-18.

Yang ZQ, Zhang ZR, Xu M, Zhu Q (2008) Study on association of melanocortin-1 receptor (MC1R) mutations with melanin trait in Chinese domestic chickens. *Research Journal of Animal Sciences* 2(2):45-49.

Yazawa R, Hirono I, Aoki T (2006) Transgenic zebrafish expressing chicken lysozyme show resistance against bacterial diseases. *Transgenic Res*. 15(3):385-91.

Ye X, Avendano S, Dekkers JC, Lamont SJ (2006) Association of twelve immune-related genes with performance of three broiler lines in two different hygiene environments. *Poult Sci*. 85(9):1555-69.

Yilmaz A, Shen S, Adelson DL, Xavier S, Zhu JJ (2005) Identification and sequence analysis of chicken Toll-like receptors. *Immunogenetics* 56:743-53.

Youn J, Hwang SH, Cho CS, Min JK, Kim WU, Park SH, Kim HY (2000) Association of the interleukin-4 receptor alpha variant Q576R with Thl/Th2 imbalance in connective tissue disease. *Immunogenetics* 51(8-9):743-6.

Zekarias B, Ter Huurne AA, Landman WJ, Rebel JM, Pol JM, Gruys E (2002) Immunological basis of differences in disease resistance in the chicken. *Vet Res*. 33(2):109-25.

Zeng K, Fu YX, Shi S, Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174(3):1431-9.

Zeng K, Mano S, Shi S, Wu CI (2007) Comparisons of site- and haplotype-frequency methods for detecting positive selection. *Mol Biol Evol*. 24(7):1562-74.

Zhai W, Nielsen R, Slatkin M (2009) An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Mol Biol Evol*. 26(2):273-83.

Zhou H, Buitenhuis AJ, Weigend S, Lamont SJ (2001) Candidate gene promoter polymorphisms and antibody response kinetics in chickens: interferon-gamma, interleukin-2, and immunoglobulin light chain. *Poult Sci.* 80(12):1679-89.

Zhou H, Gu J, Lamont SJ, Gu X (2007) Evolutionary analysis for functional divergence of the toll-like receptor gene family and altered functional constraints. *J Mol Evol.* 65:119-23.

Zhou R, Eleftheriou M, Royyuru AK, Berne BJ (2007) Destruction of long-range interactions by a single mutation in lysozyme. *Proceedings of the National Academy of Sciences of the U.S.A.* 104:5824-9.

Zongker D (2006) Chicken Chicken Chicken: Chicken Chicken. *Annals of Improbable Research* 12;16-21(6).

Appendix A – Perl scripts used

The Perl scripts are listed according to the Chapters in which they were used.

In Chapter 2

Programs used to parse Polyphred output data, ppoutParser93.pl; and to parse scanMS output: parser_scanms.pl.

ppoutParser93.pl

```
#!/var/usr/perl/
# specific to the new 93-set of Afr, As and Euros + 3 outgroups
# Program to correct genotypes and check regions in Consed ppout file and make a file of the coverage
of the reads from Consed

use warnings;
use strict;

print "\nEnter the name of your file: ";
$infile = <STDIN>;
chomp $infile;
print "Your input file is $infile\n";

print "\nEnter the name you want the output file to have: ";
$outfile = <STDIN>;
chomp $outfile;
print "Your output file is $outfile\n";

print "\nEnter the minimum numbers of reads needed for analysis (see BEGIN_COVERAGE in your ppout
$reads = <STDIN>;
chomp $reads;
print "Regions with less than $reads number of reads will not be analysed\n";

open (OUT2, ">./lyzsitestoN.txt") || die;
open (OUT3, ">./lyzCountriesTotals.txt") || die;
open (OUT4, ">./lyzCountriesFwd.txt") || die;
open (OUT5, ">./lyzCountriesRev.txt") || die;
open (OUT6, ">./lyzFreqs.txt") || die;
open (OUT7, ">./lyzRepQual.txt") || die;
open (OUT8, ">./lyzLowQualRegions.txt") || die;
open (OUT9, ">./lyztempFreqs.txt") || die;
open (OUT11, ">./temp11") || die;
open (OUT1, ">./$outfile") || die;
open (IN1, "$infile") || die "\nNo input file found";  # File with Phred-Consed data

@list = <IN1>;
$input1 = join ('', @list);
@list = split /BEGIN_CONTIG/, $input1;             # divide into each hit
# $list [0] = crap, $list[1] = Contig11 info, $list[2/3/...] = crap

@bits = split /\_/, $list[1];
# $bits[1] = polyphredranks, $bits[3] = columngenotypes, $bits[9] = manualgenotypes
# $bits[11] = verified, $bits[13] = sample, $bits[15] = coverage

@file = split /\s+/, $bits[0];
@poly = split /\n/, $bits[1];               # all ranks 1/2/3
@col = split /\n/, $bits[3];                # all genotypes, as per seq'ing
@man = split /\n/, $bits[9];                # all modified genotypes
@verified = split /\n/, $bits[11];          # listed polymorphisms (acceptable polyphred ranks 1/2/3)
@sample = split /\n/, $bits[13];            # stuff
@coverage = split /\n/, $bits[15];          # number of reads in each region
# each goes from 1 to total minus one:  [0] = BEGIN, [total] = END

$qweqwe = 0;
for($i=1; $i < scalar @verified-1 ; $i++) { # put all SNP positions in array
    @posntemp = split /\s+/, @verified[$i];
    $posn[$qweqwe] = $posntemp[0];
    $qweqwe++; }

# check coverage
@range = '';
print OUT2 "\nStartposition  Endposition  #Reads  <---- Remove this line for maskseqs3.pl";

# for loops are "-1" the length cos of "END" line as last entry in array
for($i=1; $i < scalar @coverage-1 ; $i++) {
    @cov = split /\s+/, $coverage[$i];
    # [0]/[1] = positions, [2] = #reads

    if ($cov[2] < $reads) { print OUT2 "\n$cov[0]\t\t$cov[1]\t\t$cov[2]";
```

```
        for ($r=$cov[0]; $r < $cov[1] +1 ; $r++) {  $range[$r] = $r; }    }
}
$file[3] =~ s/.REF.scf//g;
print "\n\n\t- Reading $file[3] data -\n";              # check input file is correct

@overEuroFR = '';  # free range
@overEuroBR = '';  # broiler
#etc

@overall = '';  # overall numbers of reads for each SNP in array - any
@overallF = ''; # overall numbers of reads for each SNP in array - any fwd
@overallR = ''; # overall numbers of reads for each SNP in array - any rev
#etc

@freqA = '';
#etc

# Take manual genotypes and check against column genotypes
for ($k=1; $k < scalar @col-1 ; $k++) {

    $cheq[$k] = 0;
    $poscheqx[$k] = 0;
    @column = split /\s+/, $col[$k];
    $colcheck[$k] = $col[$k];
    # [1] = position [2] = place [3] = name [4]/[5] = genotypes [6] = rank prob
    @nim = split //, $column[3];

    for ($w=0; $w < scalar @range; $w++) { if ($column[0] == $range[$w]) { $poscheqx[$k] = 1;  } }
    # ie: if column position is in regions with low #reads, don't print it later
    for($i=1; $i < scalar @man-1 ; $i++) {

        @manual = split /\s+/, $man[$i];
        # [0/1] = position [2] = place [3] = name [4] = heterozgyoteAG/homozygoteAA/indel

        if ($manual[4] =~ /heterozygote/) { $manual[4] =~ s/heterozygote//g; }
        elsif ($manual[4] =~ /homozygote/)   { $manual[4] =~ s/homozygote//g;   }
        elsif ($manual[4] eq "indel")        { $manual[4] = ''; }
                                 #  $poscheq[$k] = 1; } # ie don't print if indel
        @geno = split //, $manual[4];

        if (($manual[3] eq $column[3]) && ($manual[1] == $column[1]) && ($poscheqx[$k] == 0)) {
          $colcheck[$k] = $man[$k];
          if (!($nim[6] eq "-")) { print
OUT1"$column[0]\t$column[1]\t$column[2]\t$column[3]\t$geno[0]\t$geno[1]\t$column[6]\n";
                                     print OUT7
"#\t$column[0]\t$column[3]\t$column[4]\t$column[5]\t$column[6]\tNew =\t$geno[0]\t$geno[1]\n";
                                     if ($column[6] < 60) { print OUT8
"$column[0]\t$column[1]\t$column[2]\t$column[3]\t$geno[0]\t$geno[1]\t$column[6]\n"; } # if low
quality
       }          $cheq[$k] = 1; }    }    # change genotype

    if (($cheq[$k] == 0) && ($poscheqx[$k] == 0) && (!($nim[6] eq "-"))) { # no changed needed
        print
OUT1"$column[0]\t$column[1]\t$column[2]\t$column[3]\t$column[4]\t$column[5]\t$column[6]\n";
        if ($column[6] < 60) { print OUT8
"$column[0]\t$column[1]\t$column[2]\t$column[3]\t$column[4]\t$column[5]\t$column[6]\n"; } # if low
quality
    } }

# Check Fwd and Rev parts of column genotypes to check - bring in manual ones later
for ($k=1; $k < scalar @colcheck-1 ; $k++) {

    @column = split /\s+/, $colcheck[$k];
    if ($k > 1) { @columnPREV = split /\s+/, $colcheck[$k-1]; }
    #[0],[1] = position [2] = place [3] = name [4]/[5] = genotypes [6] = rank prob

    @sample = split //, $column[3];
    @samplePREV = split //, $columnPREV[3];
    $sam1 = $sample[5].$sample[6].$sample[7].$sample[8]; # eg 2a01, sample[9] = f/r
    $sam1PREV = $samplePREV[5].$samplePREV[6].$samplePREV[7].$samplePREV[8];

    if ($sam1 eq $sam1PREV) {
    if ((!($column[4] eq $columnPREV[4]))||(!($column[5] eq $columnPREV[5]))) { # if either don't
match
        print "\n$column[3]\t$column[4]\t$column[5]\t
$columnPREV[3]\t$columnPREV[4]\t$columnPREV[5]"; }
     # print OUT11 "\n#
$sam1\t$sam1PREV\t$column[3]\t$column[4]\t$column[5]\t$columnPREV[3]\t$columnPREV[4]\t$columnPREV[5]\
n";
    }}

# New loops to determine allele frequencies #
for ($k=1; $k < scalar @col-1 ; $k++) {
    $cheqa[$k] = 0;
    $poscheqa[$k] = 0;
    @columna = split /\s+/, $col[$k];
    # [1] = position [2] = place [3] = name [4]/[5] = genotypes [6] = rank prob
    for ($w=0; $w < scalar @range; $w++) { if ($columna[0] == $range[$w]) { $poscheqa[$k] = 1;  } }
    # ie: if column position is in regions with low #reads, don't print it later
        for($eee=0; $eee  < scalar @posn  ; $eee++) { # check each position
        if ($columna[0] == $posn[$eee]) {  # if the genotype position matches a given position
            for($i=1; $i < scalar @man-1 ; $i++) {    # need to count correct one!
                @manuala = split /\s+/, $man[$i];
                # [0/1] = position [2] = place [3] = name [4] = heterozgyoteAG/homozygoteAA/indel
               if ($manuala[4] =~ /heterozygote/) {$manuala[4] =~ s/heterozygote//g; }
                elsif ($manuala[4] =~ /homozygote/)   { $manuala[4] =~ s/homozygote//g;   }
                elsif ($manuala[4] eq "indel")        { $manuala[4] = '';
                                                    $poscheqa[$k] = 1; } # ie don't print if indel
                @geno = split //, $manuala[4];
                if (($manuala[3] eq $columna[3])&&($manuala[1] == $columna[1])&&($poscheqa[$k] == 0))
{ if   ($geno[0] eq "A" ) { $freqA[$eee]++; }
                    elsif ($geno[0] eq "C" ) { $freqC[$eee]++; }
                    elsif ($geno[0] eq "G" ) { $freqG[$eee]++; }
                    elsif ($geno[0] eq "T" ) { $freqT[$eee]++; }
                    else { $freqO[$eee]++;
   print OUT11 "$columna[0]\t$columna[1]\t$columna[2]\t$columna[3]\t$geno[0]\t$geno[1]\t$columna[6]\n";
}if    ($geno[1] eq "A" ) { $freqA[$eee]++; }
                    elsif ($geno[1] eq "C" ) { $freqC[$eee]++; }
                    elsif ($geno[1] eq "G" ) { $freqG[$eee]++; }
                    elsif ($geno[1] eq "T" ) { $freqT[$eee]++; }
                    else { $freqO[$eee]++;
  print OUT11
"$columna[0]\t$columna[1]\t$columna[2]\t$columna[3]\t$geno[0]\t$geno[1]\t$columna[6]\n"; }
                    $cheqa[$k] = 1;         }
            }   # change genotype
            if (($cheqa[$k] == 0) && ($poscheqa[$k] == 0)) { # no changed needed
                if    ($columna[4] eq "A" ) { $freqA[$eee]++; }
                elsif ($columna[4] eq "C" ) { $freqC[$eee]++; }
                elsif ($columna[4] eq "G" ) { $freqG[$eee]++; }
                elsif ($columna[4] eq "T" ) { $freqT[$eee]++; }
                else { $freqO[$eee]++;   }
                if    ($columna[5] eq "A" ) { $freqA[$eee]++; }
                elsif ($columna[5] eq "C" ) { $freqC[$eee]++; }
                elsif ($columna[5] eq "G" ) { $freqG[$eee]++; }
                elsif ($columna[5] eq "T" ) { $freqT[$eee]++; }
                else { $freqO[$eee]++; }                        } }    } }

# test each SNP pos for each sample for each country/continent/total
# Analyses regardless if reads in total are low for a region
for ($k=1; $k < scalar @col-1 ; $k++) {
    @column = split /\s+/, $col[$k];
    # [1] = position [2] = place [3] = name [4]/[5] = genotypes [6] = rank prob
    for ($w=0; $w < scalar @range; $w++) { if ($column[0] == $range[$w]) { $psch[$k] = 1;  } }
    # ie: if column position is in regions with low #reads, don't print it later

    for($eee=0; $eee  < scalar @posn  ; $eee++) {
        if (($column[0] == $posn[$eee])&&($psch[$k] != 1)) {  # if the genotype matches the position
            @letters = split //, $column[3];  # IFN-g1a01.f.ab1

            if ($letters[10] eq "f") {   # forward
                if (($letters[6] eq "a")||($letters[6] eq "d")||($letters[6] eq "f")||($letters[6] eq
"b")||($letters[6] eq "c")||($letters[6] eq "e")||($letters[6] eq "g")) { $overallF[$eee]++; }
                if    ($letters[6] eq "a") { $overfPak[$eee]++; }
                elsif ($letters[6] eq "b") { $overfBur[$eee]++; }
                elsif ($letters[6] eq "c") { $overfSen[$eee]++; }
                elsif ($letters[6] eq "d") { $overfSri[$eee]++; }
                elsif ($letters[6] eq "e") { $overfBot[$eee]++; }
                elsif ($letters[6] eq "f") { $overfBan[$eee]++; }
                elsif ($letters[6] eq "g") { $overfKen[$eee]++; }
                elsif ($letters[6] eq "h") { $overfEuroFR[$eee]++; }
                elsif ($letters[6] eq "i") { $overfEuroBR[$eee]++; }
                if (($letters[6] eq "a")||($letters[6] eq "d")||($letters[6] eq
"f")){$overfAsi[$eee]++;}
                if (($letters[6] eq "b")||($letters[6] eq "c")||($letters[6] eq "e")||($letters[6] eq
"g")){
                    $overfAfr[$eee]++;}
                if (($letters[6] eq "h")||($letters[6] eq "i")) { $overfEuro[$eee]++;}
            }

            if ($letters[10] eq "r") {  # reverse
                if (($letters[6] eq "a")||($letters[6] eq "d")||($letters[6] eq "f")||($letters[6] eq
"b")||($letters[6] eq "c")||($letters[6] eq "e")||($letter
s[6] eq "g")) { $overallR[$eee]++; }
                if    ($letters[6] eq "a") { $overrPak[$eee]++; }
                elsif ($letters[6] eq "b") { $overrBur[$eee]++; }
                elsif ($letters[6] eq "c") { $overrSen[$eee]++; }
                elsif ($letters[6] eq "d") { $overrSri[$eee]++; }
                elsif ($letters[6] eq "e") { $overrBot[$eee]++; }
                elsif ($letters[6] eq "f") { $overrBan[$eee]++; }
                elsif ($letters[6] eq "g") { $overrKen[$eee]++; }
                elsif ($letters[6] eq "h") { $overrEuroFR[$eee]++; }
                elsif ($letters[6] eq "i") { $overrEuroBR[$eee]++; }
                if(($letters[6] eq "a")||($letters[6] eq "d")||($letters[6] eq "f")) {
                    $overrAsi[$eee]++;}
                if(($letters[6] eq "b")||($letters[6] eq "c")||($letters[6] eq "e")||($letters[6] eq
"g")){
                    $overrAfr[$eee]++;}
                if(($letters[6] eq "h")||($letters[6] eq "i")) {$overrEuro[$eee]++;}}
            if    ($letters[6] eq "a") { $overPak[$eee]++; }
            elsif ($letters[6] eq "b") { $overBur[$eee]++; }
            elsif ($letters[6] eq "c") { $overSen[$eee]++; }
            elsif ($letters[6] eq "d") { $overSri[$eee]++; }
            elsif ($letters[6] eq "e") { $overBot[$eee]++; }
            elsif ($letters[6] eq "f") { $overBan[$eee]++; }
            elsif ($letters[6] eq "g") { $overKen[$eee]++; }
            elsif ($letters[6] eq "h") { $overEuroFR[$eee]++; }
            elsif ($letters[6] eq "i") { $overEuroBR[$eee]++; }

            if (($letters[6] eq "a")||($letters[6] eq "d")||($letters[6] eq "f")) {$overAsi[$eee]++;}
            if (($letters[6] eq "b")||($letters[6] eq "c")||($letters[6] eq "e")||($letters[6] eq
"g")){
                $overAfr[$eee]++;}
            if (($letters[6] eq "h")||($letters[6] eq "i")) { $overEuro[$eee]++;}

            if (($letters[6] eq "a")||($letters[6] eq "d")||($letters[6] eq "f")||($letters[6] eq
"b")||($letters[6] eq "c")||($letters[6] eq "e")||($letters[6]
 eq "g")||($letters[6] eq "h")||($letters[6] eq "i")) { $overall[$eee]++; }
        }    }}

#For SNPs 3812, 3814, 3823, 3824, 3839, 3857, 3861, 3864, 3867 there are overlapping reads;

for($u=0; $u  < scalar @posn -1 ; $u++) {
    $all[$u][0] = sprintf("%.2f", ($overall[$u]/180));   # normal
    $all[$u][1] = sprintf("%.2f", ($overAfr[$u]/80));
    $all[$u][2] = sprintf("%.2f", ($overAsi[$u]/60));
    #etc
    $all[$u][38] = sprintf("%.2f", ($overrEuroFR[$u]/10));  # EuroFR
    if (($freqA[$u]+$freqC[$u]+$freqG[$u]+$freqT[$u]) != 0) {
    $all[$u][39] = 100*(sprintf("%.3f",($freqA[$u]/($freqA[$u]+$freqC[$u]+$freqG[$u]+$freqT[$u]))));
    $all[$u][40] = 100*(sprintf("%.3f",($freqC[$u]/($freqA[$u]+$freqC[$u]+$freqG[$u]+$freqT[$u]))));
    $all[$u][41] = 100*(sprintf("%.3f",($freqG[$u]/($freqA[$u]+$freqC[$u]+$freqG[$u]+$freqT[$u]))));
    $all[$u][42] = 100*(sprintf("%.3f",($freqT[$u]/($freqA[$u]+$freqC[$u]+$freqG[$u]+$freqT[$u]))));}
    else { $zero[$u] = 1;
        print "\n$posn[$u] - zero error!"; }}

for($u=0; $u  < scalar @posn -1 ; $u++) {    # make more readable!
    for($q=0; $q  < 39 ; $q++) { $all[$u][$q] = 100*$all[$u][$q];     }       }

print OUT3"\nTOTALS\n\nSNPpos\tO/all\tAfr\tAs\t
Euro\tPakstn\tBurk_F\tSengl\tSri_L\tBotswna\tBangldh\tKenya\tBR\tFR\n";
print OUT4 "\nFWD   \n\nSNPpos\tO/all\tAfr\tAs\tE
uro\tPakstn\tBurk_F\tSengl\tSri_L\tBotswna\tBangldh\tKenya\tBR\tFR\n";
print OUT5 "\nREV\n\nSNPpos\tO/all\tAfr\tAs\t
Euro\tPakstn\tBurk_F\tSengl\tSri_L\tBotswna\tBangldh\tKenya\tBR\tFR\n";
print OUT6 "% Allele Frequencies for $infile\n\nPos\tMajor Allele\t\tMinor Alleles";

for($ree=0; $ree  < scalar @posn -1 ; $ree++) {
    if ($zero[$ree] != 1) {
        print OUT3 "\n$posn[$ree]";
        for ($w=0 ; $w < 13; $w++) { print OUT3 "\t$all[$ree][$w]"; }
        print OUT4 "\n$posn[$ree]";
        for ($w=13; $w < 26; $w++) { print OUT4 "\t$all[$ree][$w]"; }
        print OUT5 "\n$posn[$ree]";
        for ($w=26; $w < 39; $w++) { print OUT5 "\t$all[$ree][$w]"; }
        print OUT6 "\n$posn[$ree]";
        print OUT9 "\n$posn[$ree]";

        for ($w=39; $w < 43; $w++) {
   if (($all[$ree][$w] >= $all[$ree][39]) && ($all[$ree][$w] >= $all[$ree][40] ) && ($all[$ree][$w]
>= $all[$ree][41]) && ($all[$ree][$w] >= $all[$ree][42])){
                if    ($w == 39) { $best = "A"; }
                elsif ($w == 40) { $best = "C"; }
                elsif ($w == 41) { $best = "G"; }
                elsif ($w == 42) { $best = "T"; }
                print OUT6 "\t$best\t$all[$ree][$w]\t";
                for ($wr=39; $wr < 43; $wr++) {
                    if ($w != $wr) { if    ($wr == 39) { $best2 = "A"; }
                                     elsif ($wr == 40) { $best2 = "C"; }
                                     elsif ($wr == 41) { $best2 = "G"; }
                                     elsif ($wr == 42) { $best2 = "T"; }
                                     print OUT6 "\t$best2\t$all[$ree][$wr]";}     }
                if ($all[$ree][$w] <= 90.0) { print OUT6 "\t***";}
            }        }    } }

print "\nThen run \"perl PhaseIn.pl $outfile [Youroutputfilename]\" and Phase:";
print "\n\"./PHASE -d1 [Youroutputfilename] [YourPhaseoutputfilename] 100 1 100\"\n\n Then use Excel
to parse the file\nAnd use the fasta consensus sequence and the Phase output to run SeqBuild.pl on
the data before using \"hashnmask.pl\"\n\nModify the ranges in the sitestoN.txt file for input into
hashnmask.pl in the format of:\n1-98\n203-207\netc\nand remove all other charactes (ie the start
line)\n\n";
exit;
```

parser_scanms.pl

```
#!/usr/bin/perl
# Program to parse scanMS file output from MOD 1-5

open (IN1, "MOD-1.out") || die "Can't open MOD-1 file\n";
open (OUT1, ">./temp-MOD1.out");
open (IN2, "MOD-2.out") || die "Can't open MOD-2 file\n";
open (OUT2, ">./temp-MOD2.out");
open (IN3, "MOD-3.out") || die "Can't open MOD-3 file\n";
open (OUT3, ">./temp-MOD3.out");
open (IN4, "MOD-4.out") || die "Can't open MOD-4 file\n";
open (OUT4, ">./temp-MOD4.out");
open (IN5, "MOD-5.out") || die "Can't open MOD-5 file\n";
open (OUT5, ">./temp-MOD5.out");

open (OUTA, ">./temp-MODall.out");
@dpw = '';     # difference in p/w differences
@dpwsd = '';  # .... std dev
@dd = '';      # differemce in Tajima's D
@ddsd = '';   # .... std dev
@pdmax = '';  # p value max for Tajima's D
@pdmin = '';  # ...... min
@ddmin = '';  # ... in Taj D mean min value
@ddmax = '';  # ..... max value
@dfmin = '';  # ... Fu & li's D* min value
@dfmax = '';  # ..... max
@pfmin = '';  # .... p value min
@pfmax = '';  # ...... max
@dlmin = '';  # ... Fu & Li's F* min value
@dlmax = '';  # ..... max
@plmin = '';  # .... p value min
@pfmax = '';  # ..... max

# For each array elements 1-
# likewise

@MOD1 = <IN1>;
$modin1 = join ('',@MOD1);
@MOD1 = split /\n/, $modin1;
@MOD2 = <IN2>;
$modin2 = join ('',@MOD2);
@MOD2 = split /\n/, $modin2;
@MOD3 = <IN3>;
$modin3 = join ('',@MOD3);
@MOD3 = split /\n/, $modin3;
@MOD4 = <IN4>;
$modin4 = join ('',@MOD4);
@MOD4 = split /\n/, $modin4;
@MOD5 = <IN5>;
$modin5 = join ('',@MOD5);
@MOD5 = split /\n/, $modin5;

for ($i = 0; $i < scalar @MOD1; $i++) {
    @bit1 = split /\s+/, $MOD1[$i];
    if ($i==4) { print OUT1 "\n\nModel 1\nIFNG\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==8) { print OUT1 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==21) || ($i==26) || ($i==31)) { print OUT1 "\n$MOD1[$i]\t$MOD1[$i+1]"; }    }
for ($i = 0; $i < scalar @MOD1; $i++) {
    @bit1 = split /\s+/, $MOD1[$i];
    if ($i==11) { print OUT1 "\nIL1B\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==15) { print OUT1 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==23) || ($i==28) || ($i==33)) { print OUT1 "\n$MOD1[$i]\t$MOD1[$i+1]"; }            }

for ($i = 0; $i < scalar @MOD2; $i++) {
    @bit1 = split /\s+/, $MOD2[$i];
    if ($i==4) { print OUT2 "\n\nModel 2\nIFNG\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==8) { print OUT2 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==21) || ($i==26) || ($i==31)) { print OUT2 "\n$MOD1[$i]\t$MOD1[$i+1]"; }    }
for ($i = 0; $i < scalar @MOD2; $i++) {
    @bit1 = split /\s+/, $MOD2[$i];
    if ($i==11) { print OUT2 "\nIL1B\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==15) { print OUT2 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==23) || ($i==28) || ($i==33)) { print OUT2 "\n$MOD1[$i]\t$MOD1[$i+1]"; }            }

for ($i = 0; $i < scalar @MOD3; $i++) {
    @bit1 = split /\s+/, $MOD3[$i];
    if ($i==4) { print OUT3 "\n\nModel 3\nIFNG\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==8) { print OUT3 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==21) || ($i==26) || ($i==31)) { print OUT3 "\n$MOD1[$i]\t$MOD1[$i+1]"; }    }
for ($i = 0; $i < scalar @MOD3; $i++) {
    @bit1 = split /\s+/, $MOD3[$i];
    if ($i==11) { print OUT3 "\nIL1B\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==15) { print OUT3 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==23) || ($i==28) || ($i==33)) { print OUT3 "\n$MOD1[$i]\t$MOD1[$i+1]"; }            }
```

```
for ($i = 0; $i < scalar @MOD4; $i++) {
    @bit1 = split /\s+/, $MOD4[$i];
    if ($i==4) { print OUT4 "\n\nModel 4\nIFNG\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==8) { print OUT4 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==21) || ($i==26) || ($i==31)) { print OUT4 "\n$MOD1[$i]\t$MOD1[$i+1]"; }   }
for ($i = 0; $i < scalar @MOD4; $i++) {
    @bit1 = split /\s+/, $MOD4[$i];
    if ($i==11) { print OUT4 "\nIL1B\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==15) { print OUT4 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==23) || ($i==28) || ($i==33)) { print OUT4 "\n$MOD1[$i]\t$MOD1[$i+1]"; }          }

for ($i = 0; $i < scalar @MOD5; $i++) {
    @bit1 = split /\s+/, $MOD5[$i];
    if ($i==4) { print OUT5 "\n\nModel 5\nIFNG\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==8) { print OUT5 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==21) || ($i==26) || ($i==31)) { print OUT5 "\n$MOD1[$i]\t$MOD1[$i+1]"; }   }
for ($i = 0; $i < scalar @MOD5; $i++) {
    @bit1 = split /\s+/, $MOD5[$i];
    if ($i==11) { print OUT5 "\nIL1B\nP/w diff = $bit1[0]\tstd dev = $bit1[2]\t"; }
    if ($i==15) { print OUT5 "\nTaj D = $bit1[0]\tstd dev = $bit1[2]"; }
    if (($i==23) || ($i==28) || ($i==33)) { print OUT5 "\n$MOD1[$i]\t$MOD1[$i+1]"; }          }

@temp ='';
for ($i = 0; $i < scalar @MOD1; $i++) {
    @a1 = split /\s+/, $MOD1[$i];
    @a2 = split /\s+/, $MOD2[$i];
    @a3 = split /\s+/, $MOD3[$i];
    @a4 = split /\s+/, $MOD4[$i];
    @a5 = split /\s+/, $MOD5[$i];
    if ($i==4) { $pdw[0] = sprintf("%.3f", abs(($a1[0] - $a2[0])/(0.005*$a1[0]+0.005*$a2[0])));  #
p/w diff differences for infg
                 $pdw[1] = sprintf("%.3f", abs(($a1[0] - $a3[0])/(0.005*$a1[0]+0.005*$a3[0])));
                 $pdw[2] = sprintf("%.3f", abs(($a1[0] - $a4[0])/(0.005*$a1[0]+0.005*$a4[0])));
                 $pdw[3] = sprintf("%.3f", abs(($a1[0] - $a5[0])/(0.005*$a1[0]+0.005*$a5[0])));
                 $pdw[4] = sprintf("%.3f", abs(($a2[0] - $a3[0])/(0.005*$a2[0]+0.005*$a3[0])));
                 $pdw[5] = sprintf("%.3f", abs(($a2[0] - $a4[0])/(0.005*$a2[0]+0.005*$a4[0])));
                 $pdw[6] = sprintf("%.3f", abs(($a2[0] - $a5[0])/(0.005*$a2[0]+0.005*$a5[0])));
                 $pdw[7] = sprintf("%.3f", abs(($a3[0] - $a4[0])/(0.005*$a3[0]+0.005*$a4[0])));
                 $pdw[8] = sprintf("%.3f", abs(($a3[0] - $a5[0])/(0.005*$a3[0]+0.005*$a5[0])));
                 $pdw[9] = sprintf("%.3f", abs(($a4[0] - $a5[0])/(0.005*$a4[0]+0.005*$a5[0])));
                 $temp[0] =$a1[0];
                 $temp[1] =$a2[0];
                 $temp[2] =$a3[0];
                 $temp[3] =$a4[0];
                 $temp[4] =$a5[0];

                 $pdwsd[0] = sprintf("%.3f", abs((($a1[2] + $a2[2])/2)/(0.005*$a1[0]+0.005*$a2[0])));
# mean % p/w diff std dev for infg
                 $pdwsd[1] = sprintf("%.3f", abs((($a1[2] + $a3[2])/2)/(0.005*$a1[0]+0.005*$a3[0])));
                 $pdwsd[2] = sprintf("%.3f", abs((($a1[2] + $a4[2])/2)/(0.005*$a1[0]+0.005*$a4[0])));
                 $pdwsd[3] = sprintf("%.3f", abs((($a1[2] + $a5[2])/2)/(0.005*$a1[0]+0.005*$a5[0])));
                 $pdwsd[4] = sprintf("%.3f", abs((($a2[2] + $a3[2])/2)/(0.005*$a2[0]+0.005*$a3[0])));
                 $pdwsd[5] = sprintf("%.3f", abs((($a2[2] + $a4[2])/2)/(0.005*$a2[0]+0.005*$a4[0])));
                 $pdwsd[6] = sprintf("%.3f", abs((($a2[2] + $a5[2])/2)/(0.005*$a2[0]+0.005*$a5[0])));
                 $pdwsd[7] = sprintf("%.3f", abs((($a3[2] + $a4[2])/2)/(0.005*$a3[0]+0.005*$a4[0])));
                 $pdwsd[8] = sprintf("%.3f", abs((($a3[2] + $a5[2])/2)/(0.005*$a3[0]+0.005*$a5[0])));
                 $pdwsd[9] = sprintf("%.3f", abs((($a4[2] + $a5[2])/2)/(0.005*$a4[0]+0.005*$a5[0])));
  }
    if ($i==8) { $dd[0] = sprintf("%.3f", abs(($a1[0] - $a2[0])));  # Tajima's D differences for infg
                 $dd[1] = sprintf("%.3f", abs(($a1[0] - $a3[0])));
                 $dd[2] = sprintf("%.3f", abs(($a1[0] - $a4[0])));
                 $dd[3] = sprintf("%.3f", abs(($a1[0] - $a5[0])));
                 $dd[4] = sprintf("%.3f", abs(($a2[0] - $a3[0])));
                 $dd[5] = sprintf("%.3f", abs(($a2[0] - $a4[0])));
                 $dd[6] = sprintf("%.3f", abs(($a2[0] - $a5[0])));
                 $dd[7] = sprintf("%.3f", abs(($a3[0] - $a4[0])));
                 $dd[8] = sprintf("%.3f", abs(($a3[0] - $a5[0])));
                 $dd[9] = sprintf("%.3f", abs(($a4[0] - $a5[0])));
                 $temp[10] =$a1[0];
                 $temp[11] =$a2[0];
                 $temp[12] =$a3[0];
                 $temp[13] =$a4[0];
                 $temp[14] =$a5[0];

            $ddsd[0] = sprintf("%.3f", abs((($a1[2] + $a2[2])/2)));  # Mean % std dev Tajima's D  for
infg
                 $ddsd[1] = sprintf("%.3f", abs((($a1[2] + $a3[2])/2)));
                 $ddsd[2] = sprintf("%.3f", abs((($a1[2] + $a4[2])/2)));
                 $ddsd[3] = sprintf("%.3f", abs((($a1[2] + $a5[2])/2)));
                 $ddsd[4] = sprintf("%.3f", abs((($a2[2] + $a3[2])/2)));
                 $ddsd[5] = sprintf("%.3f", abs((($a2[2] + $a4[2])/2)));
                 $ddsd[6] = sprintf("%.3f", abs((($a2[2] + $a5[2])/2)));
                 $ddsd[7] = sprintf("%.3f", abs((($a3[2] + $a4[2])/2)));
                 $ddsd[8] = sprintf("%.3f", abs((($a3[2] + $a5[2])/2)));
                 $ddsd[9] = sprintf("%.3f", abs((($a4[2] + $a5[2])/2)));  }
    if ($i==11) { $pdw[10] = sprintf("%.3f", abs(($a1[0] - $a2[0])/(0.005*$a1[0]+0.005*$a2[0]))); #
p/w diff diff for il1b
                  $pdw[11] = sprintf("%.3f", abs(($a1[0] - $a3[0])/(0.005*$a1[0]+0.005*$a3[0])));
                  $pdw[12] = sprintf("%.3f", abs(($a1[0] - $a4[0])/(0.005*$a1[0]+0.005*$a4[0])));
                  $pdw[13] = sprintf("%.3f", abs(($a1[0] - $a5[0])/(0.005*$a1[0]+0.005*$a5[0])));
                  $pdw[14] = sprintf("%.3f", abs(($a2[0] - $a3[0])/(0.005*$a2[0]+0.005*$a3[0])));
                  $pdw[15] = sprintf("%.3f", abs(($a2[0] - $a4[0])/(0.005*$a2[0]+0.005*$a4[0])));
                  $pdw[16] = sprintf("%.3f", abs(($a2[0] - $a5[0])/(0.005*$a2[0]+0.005*$a5[0])));
                  $pdw[17] = sprintf("%.3f", abs(($a3[0] - $a4[0])/(0.005*$a3[0]+0.005*$a4[0])));
                  $pdw[18] = sprintf("%.3f", abs(($a3[0] - $a5[0])/(0.005*$a3[0]+0.005*$a5[0])));
                  $pdw[19] = sprintf("%.3f", abs(($a4[0] - $a5[0])/(0.005*$a4[0]+0.005*$a5[0])));
                  $temp[5] =$a1[0];
                  $temp[6] =$a2[0];
                  $temp[7] =$a3[0];
                  $temp[8] =$a4[0];
                  $temp[9] =$a5[0];

                 $pdwsd[10] = sprintf("%.3f", abs((($a1[2] + $a2[2])/2)/(0.005*$a1[0]+0.005*$a2[0])));  # p/w
diff differences for il1b
                 $pdwsd[11] = sprintf("%.3f", abs((($a1[2] + $a3[2])/2)/(0.005*$a1[0]+0.005*$a3[0])));
                 $pdwsd[12] = sprintf("%.3f", abs((($a1[2] + $a4[2])/2)/(0.005*$a1[0]+0.005*$a4[0])));
                 $pdwsd[13] = sprintf("%.3f", abs((($a1[2] + $a5[2])/2)/(0.005*$a1[0]+0.005*$a5[0])));
                 $pdwsd[14] = sprintf("%.3f", abs((($a2[2] + $a3[2])/2)/(0.005*$a2[0]+0.005*$a3[0])));
                 $pdwsd[15] = sprintf("%.3f", abs((($a2[2] + $a4[2])/2)/(0.005*$a2[0]+0.005*$a4[0])));
                 $pdwsd[16] = sprintf("%.3f", abs((($a2[2] + $a5[2])/2)/(0.005*$a2[0]+0.005*$a5[0])));
                 $pdwsd[17] = sprintf("%.3f", abs((($a3[2] + $a4[2])/2)/(0.005*$a3[0]+0.005*$a4[0])));
                 $pdwsd[18] = sprintf("%.3f", abs((($a3[2] + $a5[2])/2)/(0.005*$a3[0]+0.005*$a5[0])));
                 $pdwsd[19] = sprintf("%.3f", abs((($a4[2] + $a5[2])/2)/(0.005*$a4[0]+0.005*$a5[0]))); }

    if ($i==15) { $dd[10] = sprintf("%.3f", abs(($a1[0] - $a2[0])));  # Tajima's D differences for
il1b
                  $dd[11] = sprintf("%.3f", abs(($a1[0] - $a3[0])));
                  $dd[12] = sprintf("%.3f", abs(($a1[0] - $a4[0])));
                  $dd[13] = sprintf("%.3f", abs(($a1[0] - $a5[0])));
                  $dd[14] = sprintf("%.3f", abs(($a2[0] - $a3[0])));
                  $dd[15] = sprintf("%.3f", abs(($a2[0] - $a4[0])));
                  $dd[16] = sprintf("%.3f", abs(($a2[0] - $a5[0])));
                  $dd[17] = sprintf("%.3f", abs(($a3[0] - $a4[0])));
                  $dd[18] = sprintf("%.3f", abs(($a3[0] - $a5[0])));
                  $dd[19] = sprintf("%.3f", abs(($a4[0] - $a5[0])));
                  $temp[15] =$a1[0];
                  $temp[16] =$a2[0];
                  $temp[17] =$a3[0];
                  $temp[18] =$a4[0];
                  $temp[19] =$a5[0];
     $ddsd[10] = sprintf("%.3f", abs((($a1[2] + $a2[2])/2)));  # Mean % std dev Tajima's D  for il1b
                  $ddsd[11] = sprintf("%.3f", abs((($a1[2] + $a3[2])/2)));
                  $ddsd[12] = sprintf("%.3f", abs((($a1[2] + $a4[2])/2)));
                  $ddsd[13] = sprintf("%.3f", abs((($a1[2] + $a5[2])/2)));
                  $ddsd[14] = sprintf("%.3f", abs((($a2[2] + $a3[2])/2)));
                  $ddsd[15] = sprintf("%.3f", abs((($a2[2] + $a4[2])/2)));
                  $ddsd[16] = sprintf("%.3f", abs((($a2[2] + $a5[2])/2)));
                  $ddsd[17] = sprintf("%.3f", abs((($a3[2] + $a4[2])/2)));
                  $ddsd[18] = sprintf("%.3f", abs((($a3[2] + $a5[2])/2)));
                  $ddsd[19] = sprintf("%.3f", abs((($a4[2] + $a5[2])/2)));   }

    if ($i==22) {  $ddmin[0] = sprintf("%.3f", abs(($a1[11] - $a2[11])));  # Tajima's D differences
mean min for infg
                   $ddmin[1] = sprintf("%.3f", abs(($a1[11] - $a3[11])));
                   $ddmin[2] = sprintf("%.3f", abs(($a1[11] - $a4[11])));
                   $ddmin[3] = sprintf("%.3f", abs(($a1[11] - $a5[11])));
                   $ddmin[4] = sprintf("%.3f", abs(($a2[11] - $a3[11])));
                   $ddmin[5] = sprintf("%.3f", abs(($a2[11] - $a4[11])));
                   $ddmin[6] = sprintf("%.3f", abs(($a2[11] - $a5[11])));
                   $ddmin[7] = sprintf("%.3f", abs(($a3[11] - $a4[11])));
                   $ddmin[8] = sprintf("%.3f", abs(($a3[11] - $a5[11])));
                   $ddmin[9] = sprintf("%.3f", abs(($a4[11] - $a5[11])));

  $ddmax[0] = sprintf("%.3f", abs(($a1[14] - $a2[14])));  # Tajima's D differences mean max for infg
                   $ddmax[1] = sprintf("%.3f", abs(($a1[14] - $a3[14])));
                   $ddmax[2] = sprintf("%.3f", abs(($a1[14] - $a4[14])));
                   $ddmax[3] = sprintf("%.3f", abs(($a1[14] - $a5[14])));
                   $ddmax[4] = sprintf("%.3f", abs(($a2[14] - $a3[14])));
                   $ddmax[5] = sprintf("%.3f", abs(($a2[14] - $a4[14])));
                   $ddmax[6] = sprintf("%.3f", abs(($a2[14] - $a5[14])));
                   $ddmax[7] = sprintf("%.3f", abs(($a3[14] - $a4[14])));
                   $ddmax[8] = sprintf("%.3f", abs(($a3[14] - $a5[14])));
                   $ddmax[9] = sprintf("%.3f", abs(($a4[14] - $a5[14])));
                   $pdmin[0] = sprintf("%.3f", $a1[4]);
                   $pdmin[1] = sprintf("%.3f", $a2[4]);
                   $pdmin[2] = sprintf("%.3f", $a3[4]);
                   $pdmin[3] = sprintf("%.3f", $a4[4]);
                   $pdmin[4] = sprintf("%.3f", $a5[4]);
                   $pdmax[0] = sprintf("%.3f", $a1[7]);
                   $pdmax[1] = sprintf("%.3f", $a2[7]);
                   $pdmax[2] = sprintf("%.3f", $a3[7]);
                   $pdmax[3] = sprintf("%.3f", $a4[7]);
                   $pdmax[4] = sprintf("%.3f", $a5[7]);  }
    if ($i==27) {  $dfmin[0] = sprintf("%.3f", abs(($a1[11] - $a2[11])));  # Fu & Li's D* mean min
for infg
                     $dfmin[1] = sprintf("%.3f", abs(($a1[11] - $a3[11])));
                     $dfmin[2] = sprintf("%.3f", abs(($a1[11] - $a4[11])));
                     $dfmin[3] = sprintf("%.3f", abs(($a1[11] - $a5[11])));
                     $dfmin[4] = sprintf("%.3f", abs(($a2[11] - $a3[11])));
                     $dfmin[5] = sprintf("%.3f", abs(($a2[11] - $a4[11])));
                     $dfmin[6] = sprintf("%.3f", abs(($a2[11] - $a5[11])));
                     $dfmin[7] = sprintf("%.3f", abs(($a3[11] - $a4[11])));
                     $dfmin[8] = sprintf("%.3f", abs(($a3[11] - $a5[11])));
                     $dfmin[9] = sprintf("%.3f", abs(($a4[11] - $a5[11])));
            $dfmax[0] = sprintf("%.3f", abs(($a1[14] - $a2[14])));  # Fu & Li's D* mean max for infg
                     $dfmax[1] = sprintf("%.3f", abs(($a1[14] - $a3[14])));
                     $dfmax[2] = sprintf("%.3f", abs(($a1[14] - $a4[14])));
                     $dfmax[3] = sprintf("%.3f", abs(($a1[14] - $a5[14])));
                     $dfmax[4] = sprintf("%.3f", abs(($a2[14] - $a3[14])));
                     $dfmax[5] = sprintf("%.3f", abs(($a2[14] - $a4[14])));
                     $dfmax[6] = sprintf("%.3f", abs(($a2[14] - $a5[14])));
                     $dfmax[7] = sprintf("%.3f", abs(($a3[14] - $a4[14])));
                     $dfmax[8] = sprintf("%.3f", abs(($a3[14] - $a5[14])));
                     $dfmax[9] = sprintf("%.3f", abs(($a4[14] - $a5[14])));
                     $pfmin[0] = sprintf("%.3f", $a1[4]);
                     $pfmin[1] = sprintf("%.3f", $a2[4]);
                     $pfmin[2] = sprintf("%.3f", $a3[4]);
                     $pfmin[3] = sprintf("%.3f", $a4[4]);
                     $pfmin[4] = sprintf("%.3f", $a5[4]);
                     $pfmax[0] = sprintf("%.3f", $a1[7]);
                     $pfmax[1] = sprintf("%.3f", $a2[7]);
                     $pfmax[2] = sprintf("%.3f", $a3[7]);
                     $pfmax[3] = sprintf("%.3f", $a4[7]);
                     $pfmax[4] = sprintf("%.3f", $a5[7]);      }

    if ($i==32) {  $dlmin[0] = sprintf("%.3f", abs(($a1[11] - $a2[11])));  # Fu & Li's F* mean min
for infg
                     $dlmin[1] = sprintf("%.3f", abs(($a1[11] - $a3[11])));
                     $dlmin[2] = sprintf("%.3f", abs(($a1[11] - $a4[11])));
                     $dlmin[3] = sprintf("%.3f", abs(($a1[11] - $a5[11])));
                     $dlmin[4] = sprintf("%.3f", abs(($a2[11] - $a3[11])));
                     $dlmin[5] = sprintf("%.3f", abs(($a2[11] - $a4[11])));
                     $dlmin[6] = sprintf("%.3f", abs(($a2[11] - $a5[11])));
                     $dlmin[7] = sprintf("%.3f", abs(($a3[11] - $a4[11])));
                     $dlmin[8] = sprintf("%.3f", abs(($a3[11] - $a5[11])));
                     $dlmin[9] = sprintf("%.3f", abs(($a4[11] - $a5[11])));
          $dlmax[0] = sprintf("%.3f", abs(($a1[14] - $a2[14])));  # Fu & Li's F* mean max for infg
                     $dlmax[1] = sprintf("%.3f", abs(($a1[14] - $a3[14])));
                     $dlmax[2] = sprintf("%.3f", abs(($a1[14] - $a4[14])));
                     $dlmax[3] = sprintf("%.3f", abs(($a1[14] - $a5[14])));
                     $dlmax[4] = sprintf("%.3f", abs(($a2[14] - $a3[14])));
                     $dlmax[5] = sprintf("%.3f", abs(($a2[14] - $a4[14])));
                     $dlmax[6] = sprintf("%.3f", abs(($a2[14] - $a5[14])));
                     $dlmax[7] = sprintf("%.3f", abs(($a3[14] - $a4[14])));
                     $dlmax[8] = sprintf("%.3f", abs(($a3[14] - $a5[14])));
                     $dlmax[9] = sprintf("%.3f", abs(($a4[14] - $a5[14])));
                     $plmin[0] = sprintf("%.3f", $a1[4]);
                     $plmin[1] = sprintf("%.3f", $a2[4]);
                     $plmin[2] = sprintf("%.3f", $a3[4]);
                     $plmin[3] = sprintf("%.3f", $a4[4]);
                     $plmin[4] = sprintf("%.3f", $a5[4]);
                     $plmax[0] = sprintf("%.3f", $a1[7]);
                     $plmax[1] = sprintf("%.3f", $a2[7]);
                     $plmax[2] = sprintf("%.3f", $a3[7]);
                     $plmax[3] = sprintf("%.3f", $a4[7]);
                     $plmax[4] = sprintf("%.3f", $a5[7]);     }

################## I L 1 B ##################

    if ($i==24) {  $ddmin[10] = sprintf("%.3f", abs(($a1[11] - $a2[11])));  # Tajima's D differences
mean min for il1b
                     $ddmin[11] = sprintf("%.3f", abs(($a1[11] - $a3[11])));
                     $ddmin[12] = sprintf("%.3f", abs(($a1[11] - $a4[11])));
                     $ddmin[13] = sprintf("%.3f", abs(($a1[11] - $a5[11])));
                     $ddmin[14] = sprintf("%.3f", abs(($a2[11] - $a3[11])));
                     $ddmin[15] = sprintf("%.3f", abs(($a2[11] - $a4[11])));
                     $ddmin[16] = sprintf("%.3f", abs(($a2[11] - $a5[11])));
                     $ddmin[17] = sprintf("%.3f", abs(($a3[11] - $a4[11])));
                     $ddmin[18] = sprintf("%.3f", abs(($a3[11] - $a5[11])));
                     $ddmin[19] = sprintf("%.3f", abs(($a4[11] - $a5[11])));
 $ddmax[10] = sprintf("%.3f", abs(($a1[14] - $a2[14])));  # Tajima's D differences mean max for il1b
                     $ddmax[11] = sprintf("%.3f", abs(($a1[14] - $a3[14])));
                     $ddmax[12] = sprintf("%.3f", abs(($a1[14] - $a4[14])));
                     $ddmax[13] = sprintf("%.3f", abs(($a1[14] - $a5[14])));
                     $ddmax[14] = sprintf("%.3f", abs(($a2[14] - $a3[14])));
                     $ddmax[15] = sprintf("%.3f", abs(($a2[14] - $a4[14])));
                     $ddmax[16] = sprintf("%.3f", abs(($a2[14] - $a5[14])));
                     $ddmax[17] = sprintf("%.3f", abs(($a3[14] - $a4[14])));
                     $ddmax[18] = sprintf("%.3f", abs(($a3[14] - $a5[14])));
                     $ddmax[19] = sprintf("%.3f", abs(($a4[14] - $a5[14])));
```

```
                    $pdmin[5] = sprintf("%.3f", $a1[4]);
                    $pdmin[6] = sprintf("%.3f", $a2[4]);
                    $pdmin[7] = sprintf("%.3f", $a3[4]);
                    $pdmin[8] = sprintf("%.3f", $a4[4]);
                    $pdmin[9] = sprintf("%.3f", $a5[4]);
                    $pdmax[5] = sprintf("%.3f", $a1[7]);
                    $pdmax[6] = sprintf("%.3f", $a2[7]);
                    $pdmax[7] = sprintf("%.3f", $a3[7]);
                    $pdmax[8] = sprintf("%.3f", $a4[7]);
                    $pdmax[9] = sprintf("%.3f", $a5[7]);    }

    if ($i==29) {   $dfmin[10] = sprintf("%.3f", abs(($a1[11] - $a2[11])));  # Fu & Li's D* mean min
for il1b
                    $dfmin[11] = sprintf("%.3f", abs(($a1[11] - $a3[11])));
                    $dfmin[12] = sprintf("%.3f", abs(($a1[11] - $a4[11])));
                    $dfmin[13] = sprintf("%.3f", abs(($a1[11] - $a5[11])));
                    $dfmin[14] = sprintf("%.3f", abs(($a2[11] - $a3[11])));
                    $dfmin[15] = sprintf("%.3f", abs(($a2[11] - $a4[11])));
                    $dfmin[16] = sprintf("%.3f", abs(($a2[11] - $a5[11])));
                    $dfmin[17] = sprintf("%.3f", abs(($a3[11] - $a4[11])));
                    $dfmin[18] = sprintf("%.3f", abs(($a3[11] - $a5[11])));
                    $dfmin[19] = sprintf("%.3f", abs(($a4[11] - $a5[11])));
          $dfmax[10] = sprintf("%.3f", abs(($a1[14] - $a2[14])));  # Fu & Li's D* mean max for il1b
                    $dfmax[11] = sprintf("%.3f", abs(($a1[14] - $a3[14])));
                    $dfmax[12] = sprintf("%.3f", abs(($a1[14] - $a4[14])));
                    $dfmax[13] = sprintf("%.3f", abs(($a1[14] - $a5[14])));
                    $dfmax[14] = sprintf("%.3f", abs(($a2[14] - $a3[14])));
                    $dfmax[15] = sprintf("%.3f", abs(($a2[14] - $a4[14])));
                    $dfmax[16] = sprintf("%.3f", abs(($a2[14] - $a5[14])));
                    $dfmax[17] = sprintf("%.3f", abs(($a3[14] - $a4[14])));
                    $dfmax[18] = sprintf("%.3f", abs(($a3[14] - $a5[14])));
                    $dfmax[19] = sprintf("%.3f", abs(($a4[14] - $a5[14])));
                    $pfmin[5] = sprintf("%.3f", $a1[4]);
                    $pfmin[6] = sprintf("%.3f", $a2[4]);
                    $pfmin[7] = sprintf("%.3f", $a3[4]);
                    $pfmin[8] = sprintf("%.3f", $a4[4]);
                    $pfmin[9] = sprintf("%.3f", $a5[4]);
                    $pfmax[5] = sprintf("%.3f", $a1[7]);
                    $pfmax[6] = sprintf("%.3f", $a2[7]);
                    $pfmax[7] = sprintf("%.3f", $a3[7]);
                    $pfmax[8] = sprintf("%.3f", $a4[7]);
                    $pfmax[9] = sprintf("%.3f", $a5[7]);    }

    if ($i==34) {   $dlmin[10] = sprintf("%.3f", abs(($a1[11] - $a2[11])));  # Fu & Li's F* mean min
for il1b
                    $dlmin[11] = sprintf("%.3f", abs(($a1[11] - $a3[11])));
                    $dlmin[12] = sprintf("%.3f", abs(($a1[11] - $a4[11])));
                    $dlmin[13] = sprintf("%.3f", abs(($a1[11] - $a5[11])));
                    $dlmin[14] = sprintf("%.3f", abs(($a2[11] - $a3[11])));
                    $dlmin[15] = sprintf("%.3f", abs(($a2[11] - $a4[11])));
                    $dlmin[16] = sprintf("%.3f", abs(($a2[11] - $a5[11])));
                    $dlmin[17] = sprintf("%.3f", abs(($a3[11] - $a4[11])));
                    $dlmin[18] = sprintf("%.3f", abs(($a3[11] - $a5[11])));
                    $dlmin[19] = sprintf("%.3f", abs(($a4[11] - $a5[11])));
            $dlmax[10] = sprintf("%.3f", abs(($a1[14] - $a2[14])));  # Fu & Li's F* mean max for il1b
                    $dlmax[11] = sprintf("%.3f", abs(($a1[14] - $a3[14])));
                    $dlmax[12] = sprintf("%.3f", abs(($a1[14] - $a4[14])));
                    $dlmax[13] = sprintf("%.3f", abs(($a1[14] - $a5[14])));
                    $dlmax[14] = sprintf("%.3f", abs(($a2[14] - $a3[14])));
                    $dlmax[15] = sprintf("%.3f", abs(($a2[14] - $a4[14])));
                    $dlmax[16] = sprintf("%.3f", abs(($a2[14] - $a5[14])));
                    $dlmax[17] = sprintf("%.3f", abs(($a3[14] - $a4[14])));
                    $dlmax[18] = sprintf("%.3f", abs(($a3[14] - $a5[14])));
                    $dlmax[19] = sprintf("%.3f", abs(($a4[14] - $a5[14])));
                    $plmin[5] = sprintf("%.3f", $a1[4]);
                    $plmin[6] = sprintf("%.3f", $a2[4]);
                    $plmin[7] = sprintf("%.3f", $a3[4]);
                    $plmin[8] = sprintf("%.3f", $a4[4]);
                    $plmin[9] = sprintf("%.3f", $a5[4]);
                    $plmax[5] = sprintf("%.3f", $a1[7]);
                    $plmax[6] = sprintf("%.3f", $a2[7]);
                    $plmax[7] = sprintf("%.3f", $a3[7]);
                    $plmax[8] = sprintf("%.3f", $a4[7]);
                    $plmax[9] = sprintf("%.3f", $a5[7]);    }
}

print "\nModel\t\t\t\tP/w Differences:\tinfg\t\til1b\tTajima's D:\tinfg\til1b\n1\tDomestication:
exponential growth after stasis\t$temp[0]\t$temp[5]\t$temp[10]\
t$temp[15]\n2\tIL1B: rise & rise\t\t\t\t$temp[1]\t$temp[6]\t$temp[11]\t$temp[16]\n3\tIFN-g: rise,
stasis & fall\t\t\t$temp[2]\t$temp[7]\t$temp[12]\t$temp[17]\n
4\tSteady constant increase\t\t\t$temp[3]\t$temp[8]\t$temp[13]\t$temp[18]\n5\tPermanent
stasis\t\t\t\t$temp[4]\t$temp[9]\t$temp[14]\t$temp[19]\n";

for ($x=0; $x < scalar @pdw; $x++) {      if ($x == 0) {print "\nINFG % p/w diff differences b/w models"; }
                                          if ($x ==10) {print "\nIL1B % p/w diff differences b/w models"; }
                                          if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                          print "$pdw[$x]\t";                              }
```

```
for ($x=0; $x < scalar @pdwsd; $x++) {  if ($x == 0) {print "\nINFG average % p/w diff std dev (% of total
p/w diff)"; }
                                         if ($x ==10) {print "\nIL1B average % p/w diff std dev (% of total

p/w diff)"; }
                                         if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                         print "$pdwsd[$x]\t";                                          }
for ($x=0; $x < scalar @dd; $x++) {   if ($x == 0) {print "\nINFG Tajima's D differences"; }
                                      if ($x ==10) {print "\nIL1B Tajima's D differences"; }
                                      if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                      print "$dd[$x]\t";                                                }
for ($x=0; $x < scalar @ddsd; $x++) { if ($x == 0) {print "\nINFG average Tajima's D std dev (mean of
models)"; }
                                      if ($x ==10) {print "\nIL1B average Tajima's D std dev (mean of

models)"; }
                                      if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                      print "$ddsd[$x]\t";                                        }
for ($x=0; $x < scalar @ddmin; $x++) { if ($x == 0) {print "\nINFG Tajima's D differences mean min"; }
                                       if ($x ==10) {print "\nIL1B Tajima's D differences mean min"; }
                                       if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                       print "$ddmin[$x]\t";                                      }
for ($x=0; $x < scalar @ddmax; $x++) { if ($x == 0) {print "\nINFG Tajima's D differnces mean max"; }
                                       if ($x ==10) {print "\nIL1B Tajima's D differences mean max"; }
                                       if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                       print "$ddmax[$x]\t";                                      }
for ($x=0; $x < scalar @dfmin; $x++) { if ($x == 0) {print "\nINFG Fu & Li's D* differnces mean min"; }
                                       if ($x ==10) {print "\nIL1B Fu & Li's D* differences mean min"; }
                                       if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                       print "$dfmin[$x]\t";                                      }
for ($x=0; $x < scalar @dfmax; $x++) { if ($x == 0) {print "\nINFG Fu & Li's D* differences mean max"; }
                                       if ($x ==10) {print "\nIL1B Fu & Li's D* differences mean max"; }
                                       if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                       print "$dfmax[$x]\t";                                       }
for ($x=0; $x < scalar @dlmin; $x++) { if ($x == 0) {print "\nINFG average % Fu & Li's F* mean min"; }
                                       if ($x ==10) {print "\nIL1B average % Fu & Li's F* mean min"; }
                                       if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                       print "$dlmin[$x]\t";                                       }
for ($x=0; $x < scalar @dlmax; $x++) { if ($x == 0) {print "\nINFG Fu & Li's F* differences mean max"; }
                                       if ($x ==10) {print "\nIL1B Fu & Li's F* differences mean max"; }
                                       if ($x==0 || $x ==10) { print "\nModels:\t1-2\t1-3\t1-4\t1-5\t2-3\t2-
4\t2-5\t3-4\t3-5\t4-5\n\t"; }
                                       print "$dlmax[$x]\t";                                       }
print "\n\n\tP Values";
for ($x=0; $x < scalar @pdmin; $x++) { if ($x==0) { print "\nIL1B Tajima's D min\nModel\t1\t2\t3\t4\t5\n\t";}
                                       if ($x==5) { print "\nIFNG Tajima's D min\nModel\t1\t2\t3\t4\t5\n\t";}
                                       print "$pdmin[$x]\t";
                                     }
for ($x=0; $x < scalar @pdmax; $x++) { if ($x==0) { print "\nIL1B Tajima's D max\nModel\t1\t2\t3\t4\t5\n\t";}
                                       if ($x==5) { print "\nIFNG Tajima's D max\nModel\t1\t2\t3\t4\t5\n\t";}
                                       print "$pdmax[$x]\t";                                       }
for ($x=0; $x < scalar @pfmin; $x++) { if ($x==0) { print "\nIL1B Fu & Li's D*
min\nModel\t1\t2\t3\t4\t5\n\t"; }
                                       if ($x==5) { print "\nIFNG Fu & Li's D*

min\nModel\t1\t2\t3\t4\t5\n\t"; }
                                       print "$pfmin[$x]\t";                                       }
for ($x=0; $x < scalar @pfmax; $x++) { if ($x==0) { print "\nIL1B Fu & Li's D*
max\nModel\t1\t2\t3\t4\t5\n\t"; }
                                       if ($x==5) { print "\nIFNG Fu & Li's D*

max\nModel\t1\t2\t3\t4\t5\n\t"; }
                                       print "$pfmax[$x]\t";                                       }
for ($x=0; $x < scalar @plmin; $x++) { if ($x==0) { print "\nIL1B Fu & Li's F*
min\nModel\t1\t2\t3\t4\t5\n\t"; }
                                       if ($x==5) { print "\nIFNG Fu & Li's F*

min\nModel\t1\t2\t3\t4\t5\n\t"; }
                                       print "$plmin[$x]\t";                                       }
for ($x=0; $x < scalar @plmax; $x++) { if ($x==0) { print "\nIL1B Fu & Li's F*
max\nModel\t1\t2\t3\t4\t5\n\t"; }
                                       if ($x==5) { print "\nIFNG Fu & Li's F*

max\nModel\t1\t2\t3\t4\t5\n\t"; }
                                       print "$plmax[$x]\t";                                       }
print "\n";
exit;
```

Chapter 3

Perl scripts used in parsing and analysis of pairwise comparison of chicken with zebra finch sequences: ckzfNEWblastx.pl, BLASTX.pl, hitParserZF.pl.

ckzfNEWblastx.pl

```
#!/usr/bin/perl
# Program to analyse Blastx output file - CK vs ZF contigs
# Need to incorporate separate HSPs

use DBI;
$DBD      = 'mysql';
$host     = 'popgen.gen.tcd.ie';
$user     = 'downingt';
$password = '_______';
$database = 'tim';
$dbh = DBI->connect("DBI:$DBD:$database:$host","$user","$password", { RaiseError => 1, AutoCommit =>
1});

open (OUT1, ">./temp1");
open (OUT2, ">./temp2");
open (OUT3, ">./temp3");
open (OUT4, ">./temp4");
open (OUT5, ">./temp5");
open (OUT6, ">./temp6");
open (OUT7, ">./temp7");
open (OUT8, ">./temp8");
open (OUT9, ">./temp9");
open (OUT10, ">./temp10");
open (IN, "CKrsmrnaVCKestcontigsBLASTN.out") || die;

$t = 0;
@theframes = '';
@zfcheck = '';
@ckcheck = '';
@blastx = <IN>;
$input1 = join ('', @blastx);
@blastx = split /Reference/, $input1;              # divide into each hit
print OUT10 "Total number of entries = ", scalar @blastx;

for ($i=1; $i < scalar @blastx; $i++) {   # $i = hit, $j = hsp

    @hsp = split /Score =/, $blastx[$i];      # hsp[0] = crap, [1] = first hsp, [2] = second etc ...
    @hspone = split /Query=/, $hsp[0];            # get names of CK/ZF etc
    @zf = split /\s+/, $hspone[1];
    $zf[2] =~ s/\(//g;                            # ZF length
    @hsptwo = split /Value/, $hsp[0];
    @ck = split /\s+/, $hsptwo[1];                # CK length, e-value etc

    for ($j=1; $j < scalar @hsp; $j++) {# for each HSP of each hit (skip zero, != hsp)

        @det = split /\s+/, $hsp[$j];
        $det[10] =~ s/\(//g;
        $det[10] =~ s/\)//g;
        $det[10] =~ s/\,//g;
        $det[10] =~ s/\%//g;                      # get % ID
        $idpercent[$i][$j] = $det[10];

        @idl = split /\//, $det[9];               # get ID length
        $idlength[$i][$j] = $idl[1];

        @frame = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0);  # set all frames to zilch
        @frameline = split /Frame/, $hsp[$j];
        @frameC = split /\s+/, $frameline[1];
        if ($frame[$j] == 0) { $frame[$j] = $frameC[2]; }
        $theframes[$i][$j] = $frame[$j];

        if ($j > 1) {           # frame difference

if((($theframes[$i][$j]>0)&&($theframes[$i][1]>0))||(($theframes[$i][$j]<0)&&($theframes[$i][1]<0))){
            $frameDiff[$i][$j] = $theframes[$i][$j] - $theframes[$i][1]; }
            else { $idlength[$i][$j] = 0; }
        }
        elsif ($j == 1) { $frameDiff[$i][1] = 0; }

        if (($idlength[$i][$j] > 70) && ($idpercent[$i][$j] > 60)) {
            # if ID% > 60%, if ID length > 65
            #print OUT1 "\n$zf[1]\t$ck[1]\tZFlength = $zf[2]\tCKlength = $ck[7]";
            #print OUT1 "\tScore = $ck[2]\nEval = $ck[3]\tID% = $det[10]%\tIDlength = $idl[1]";
            #print OUT1 "\tFrame=\t$theframes[$i][$j]\tHit $i\tHSP $j";
            @zfdets = split /Query:/, $hsp[$j]; # ZF
            $zfprot = '';
            $zfstart = -1;
            $zfend = -1;
            @ckdets = split /Sbjct:/, $hsp[$j]; #  Ck
            $ckprot = '';
            $ckstart = -2;
            $ckend = -2;
            for ($y=1; $y < scalar @zfdets; $y++) {         # ZF
                @zfparts = split /\s+/, $zfdets[$y];
                $zfprot = $zfprot.$zfparts[2];
                $zfprot =~ s/-//g;
                if ($zfstart == -1) { $zfstart = $zfparts[1]; }
                if ($zfparts[3] > $zfend) { $zfend = $zfparts[3]; }
                if ($zfparts[3] < $zfend) { $zfendstemp = $zfparts[3]; }
            }
            for ($y=1; $y < scalar @ckdets; $y++) {        # CK
                @ckparts = split /\s+/, $ckdets[$y];
                $ckprot = $ckprot.$ckparts[2];
                $ckprot =~ s/-//g;
                if ($ckstart == -2) { $ckstart = $ckparts[1]; }
                if ($ckparts[3] > $ckend) { $ckend = $ckparts[3]; }
                if ($ckparts[3] < $ckend) { $ckendstemp = $ckparts[3]; }
            }

            $zfprotseq[$i][$j] = $zfprot;
            $ckprotseq[$i][$j] = $ckprot;
            $zfstarts[$i][$j] = $zfstart;
            $zfends[$i][$j] = $zfend;
            $ckstarts[$i][$j] = $ckstart;
            $ckends[$i][$j] = $ckend;
            $zfnames[$i] = $zf[1];
            $cknames[$i] = $ck[1];
            $zfendstemp[$i][$j] = $zfendstemp;
            $ckendstemp[$i][$j] = $ckendstemp;

            if ($j > 1) {
           #print OUT2 "\nHit $i\t$zfnames[$i]\tHSP
         $j\tFrame=$theframes[$i][$j]\tFD=$frameDiff[$i][$j]";
    #print OUT2 "\nZF start = $zfstarts[$i][$j]\tend = $zfends[$i][$j]\n$zfprotseq[$i][$j]";
    #print OUT3 "\nHit $i\t$cknames[$i]\tHSP $j\tFrame=$theframes[$i][$j]\tFD=$frameDiff[$i][$j]";
    #print OUT3 "\nCK start = $ckstarts[$i][$j]\tend = $ckends[$i][$j]\n$ckprotseq[$i][$j]";
            }

            $stw="select cknew18kdna.ckname, cknew18kdna.ckdnaseq from cknew18kdna where
cknew18kdna.ckname = '$cknames[$i]';";
            $sth = $dbh->prepare($stw) or die "Can't prepare $stw: $dbh->errstr\n";
            $rv = $sth->execute() or die "Can't execute the query: $sth->errstr";

            if ($rv eq 1)   {
                while(@array=$sth->fetchrow_array)  {
                    $chickenname = $array[0];
                    $chickenseq = $array[1];
                    $chicken[$i][0] = $chickenname;
                    $chicken[$i][$j] = $chickenseq;      }
            }

            $stw="select zfnew8kcontigs.zfname, zfnew8kcontigs.zfdnaseq from zfnew8kcontigs where
zfnew8kcontigs.zfname = '$zfnames[$i]';";
            $sth = $dbh->prepare($stw) or die "Can't prepare $stw: $dbh->errstr\n";
            $rv = $sth->execute() or die "Can't execute the query: $sth->errstr";

            if ($rv eq 1)   {
                while(@array=$sth->fetchrow_array)  {
                    $zebname = $array[0];
                    $zebseq = $array[1];
                    $zebrafinch[$i][0] = $zebname;
                    $zebrafinch[$i][$j] = $zebseq;       }
            }

            $cklong = length $chicken[$i][$j];        # lengths
            $zflong = length $zebrafinch[$i][$j];

            if ($theframes[$i][$j] < 0) {                      # if negative
                $zfends[$i][$j] = $zflong - $zfendstemp[$i][$j] + 1;
                $zfstarts[$i][$j] = $zflong - $zfstarts[$i][$j] +1;
                $zebrafinch[$i][$j] = reverse ($zebrafinch[$i][$j]);
                $zebrafinch[$i][$j] =~ tr/ACGT/TGCA/;# switch strand, starts, ends etc
            }
            if (($theframes[$i][$j] < 0) && ($ckstart[$i][$j] > $ckend[$i][$j])) {
                $ckends[$i][$j] = $cklong/3 - $ckendstemp[$i][$j] +1/3;
                $ckstarts[$i][$j] = $cklong/3 - $ckstarts[$i][$j] +1/3;
                $chicken[$i][$j] = reverse ($chicken[$i][$j]);
                $chicken[$i][$j] =~ tr/ACGT/TGCA/;
            }

            if ($zfstarts[$i][$j] < 1 ) { $zfstarts[$i][$j] == 0;}
            else { $zfstarts[$i][$j]--; }
            $ckstarts[$i][$j] = 3*$ckstarts[$i][$j];
            $ckends[$i][$j] = 3*$ckends[$i][$j];
            if ($ckstarts[$i][$j] < 3 ) { $ckstarts[$i][$j] == 0;}
            else { $ckstarts[$i][$j] -= 3; }        # -1 ????

            $ckprotseq[$i][$j] =~ s/\*//g;
            $ckprotseq[$i][$j] =~ s/-//g;
            $zfprotseq[$i][$j] =~ s/-//g;
            $zfprotseq[$i][$j] =~ s/\*//g;
            $zfprotseq[$i][$j] =~ s/[BJOUx]//g;
            $ckprotseq[$i][$j] =~ s/[BJOUx]//g;

            $ck[$i][$j] = $chicken[$i][$j];
            $zf[$i][$j] = $zebrafinch[$i][$j];

             $zeb[$i][$j]=substr($zf[$i][$j],$zfstarts[$i][$j]+1+$frameDiff[$i][$j],$zfends[$i][$j]);
             $chick[$i][$j]=substr($ck[$i][$j],$ckstarts[$i][$j]+1+$frameDiff[$i][$j],$ckends[$i][$j]);
             $zebrafinch[$i][$j]=substr($zf[$i][$j],$zfstarts[$i][$j]+$frameDiff[$i][$j],$zfends[$i][$j]);
             $chicken[$i][$j]=substr($ck[$i][$j],$ckstarts[$i][$j]+$frameDiff[$i][$j],$ckends[$i][$j]);
            if (($chicken[$i][0] =~ $cknames[$i]) && ($zebrafinch[$i][0] =~ $zfnames[$i])) {
       #print OUT4 "\n>$chicken[$i][0]\n$chicken[$i][$j]\n>$zebrafinch[$i][0]\n$zebrafinch[$i][$j]"; #DNA
       #print OUT5 "\n>$cknames[$i]\n$ckprotseq[$i][$j]\n$zfnames[$i]\n$zfprotseq[$i][$j]";      # protein
                open (OUTP, ">./prot$i-$j.fa");
                print OUTP ">$cknames[$i]\n$ckprotseq[$i][$j]\n>$zfnames[$i]\n$zfprotseq[$i][$j]";
                open (OUTD, ">./dna$i-$j.fa");
             print OUTD ">$chicken[$i][0]\n$chicken[$i][$j]\n>$zebrafinch[$i][0]\n$zebrafinch[$i][$j]\n";

                system ("transeq dna$i-$j.fa transeq$i-$j.fa -auto"); # TRANSEQ; names have "_1"
attached
                open (IN7, "transeq$i-$j.fa") || die;

                @seq = <IN7>;
                $input7 = join ('', @seq);
                $input7 =~ s/\s+//g;
                $input7 =~ s/>/\n/g;
                $input7 =~ s/_1/_1\n/g;
                $input7 =~ s/_1//g; # removes from zf only
                $input7 =~ s/\*//g;
                $input7 =~ s/[BJOUx]//g;
                @seq = split /\n/, $input7;    # zf protein = $seq[4], ck protein = $seq[2]

                if ($seq[4] =~ $zfprotseq[$i][$j]) {  $ernie++;
                                                      $zfcheck[$i][$j] = 0; }   # ZF
                else { $igor++;
                       $zfcheck[$i][$j] = 1; }

                if ($seq[2] =~ $ckprotseq[$i][$j]) { $bert++;
                                                     $ckcheck[$i][$j] = 0;} # CK
                else { $duck++;
                       $ckcheck[$i][$j] = 1;}

                for ($g=-3; $g < 3; $g++) {

$zeb[$i][$j]=substr($zf[$i][$j],$zfstarts[$i][$j]+$g+$frameDiff[$i][$j],$zfends[$i][$j]);

$chick[$i][$j]=substr($ck[$i][$j],$ckstarts[$i][$j]+$g+$frameDiff[$i][$j],$ckends[$i][$j]);

                if ($zfcheck[$i][$j] == 1) {

                    open (OUTD, ">./dna$i-$j.fa");
                    print OUTD
">$chicken[$i][0]\n$chicken[$i][$j]\n>$zebrafinch[$i][0]\n$zeb[$i][$j]\n";

                    system ("transeq dna$i-$j.fa transeq$i-$j.fa -auto");
                    open (IN8, "transeq$i-$j.fa") || die;

                    @seq2 = <IN8>;
                    $input8 = join ('', @seq2);
                    $input8 =~ s/\s+//g;
                    $input8 =~ s/>/\n/g;
                    $input8 =~ s/_1/_1\n/g;
                    $input8 =~ s/_1//g; # removes from zf only
                    $input8 =~ s/\*//g;
                    $input8 =~ s/[BJOUx]//g;
                    @seq2 = split /\n/, $input8;    # zf protein = $seq[4], ck protein = $seq[2]

                    if ($seq2[4] eq $zfprotseq[$i][$j]) {  $ernie++;
                                                           $igor--;
                                                           $zfcheck[$i][$j] = 0; }   # ZF
                    else { $zfcheck[$i][$j] = 1;}
                } # end zf check

                if ($ckcheck[$i][$j] == 1) {
                    open (OUTD, ">./dna$i-$j.fa");
               print OUTD ">$chicken[$i][0]\n$chick[$i][$j]\n>$zebrafinch[$i][0]\n$zebrafinch[$i][$j]\n";
                    system ("transeq dna$i-$j.fa transeq$i-$j.fa -auto");
                    open (IN9, "transeq$i-$j.fa") || die;
                    @seq3 = <IN9>;
                    $input9 = join ('', @seq3);
                    $input9 =~ s/\s+//g;
                    $input9 =~ s/>/\n/g;
                    $input9 =~ s/_1/_1\n/g;
                    $input9 =~ s/_1//g; # removes from zf only
                    $input9 =~ s/\*//g;
                    $input9 =~ s/[BJOUx]//g;
                    @seq3 = split /\n/, $input9;

                    if ($seq3[2] eq $ckprotseq[$i][$j]) { $bert++;
                                                          $duck--;
                                                          $ckcheck[$i][$j] = 0;} # CK
                    else { $ckcheck[$i][$j] = 1;}
                } # end ck check
            } # end $g loop

                if ($zfcheck[$i][$j] == 1) { $a++;
                     print OUT8 "\n$seq2[3]\ton $seq2[1]\tHit $i\tHSP $j\tstrand=$theframes[$i][$j]";
                     print OUT8 "\tframeDiff=$frameDiff[$i][$j]\tzfst=$zfstarts[$i][$j]";
                     print OUT8 "\tend=$zfends[$i][$j]\tL=$zflong\nTrans=$seq2[4]";
                     print OUT8 "\nReal= $zfprotseq[$i][$j]\nDNA=$zeb[$i][$j]\n______"; }

                if ($chcheck[$i][$j] == 1) { $b++;
                     print OUT9 "\n$seq2[1]\ton $seq2[3]\tHit $i\tHSP $j\tstrand=$theframes[$i][$j]";
                     print OUT9 "\tFrameDiff=$frameDiff[$i][$j]\tckst=$ckstarts[$i][$j]";
                     print OUT9 "\tend=$ckends[$i][$j]\tL=$cklong";
                     print OUT9 "\nTrans=$seq2[2]\nReal= $ckprotseq[$i][$j]\nDNA=$chick[$i][$j]\n___"
                     }
                if ($chcheck[$i][$j] == 0) { $c++; }
                if ($zfcheck[$i][$j] == 0) { $d++; }

           system ("t_coffee -infile=prot$i-$j.fa -outfile tcpt$i-$j.aln"); # T-COFFEE

                if ($chicken[$i][0] =~ /NM_001030626/) {
         print OUT10 "\n1\tNM_001030626\thit $i\thsp $j\t$t\t>$chicken[$i][0] $zebrafinch[$i][0]\n";
}

                system ("mv dna$i-$j.fa dna$t.fa");
                system ("mv prot$i-$j.fa prot$t.fa");
                system ("mv tcpt$i-$j.aln tcpt$t.aln");
                $t++;
              } # end name check
        } # end ID length, %
    }
}

print "\nTotal number = ", $i + $j -1, "\thit = $i\thsp = $j\tt = $t";
print "\nWorking zf = $ernie\tck = $bert\t not zf = $igor\t not ck = $duck\n";
print "****not okay for zf = $a\tck = $b\tokay zf = $d\tck = $c\n";

exit;
```

BLASTX.pl

```
#!/usr/bin/perl
# Program to analyse the Blastx output file

use DBI;
$DBD      = 'mysql';
$host     = 'popgen.gen.tcd.ie';
$user     = 'downingt';
$password = '_______';
$database = 'tim';
$dbh = DBI->connect("DBI:$DBD:$database:$host", "$user", "$password", { RaiseError => 1, AutoCommit
=> 1});

@zftemp1 = '';
@cktemp1 = '';
$count1 = 0;
$count2 = 0;
@array1 = '';
@array2 = '';
@array3 = '';
@array4 = '';
@array5 = '';
@array6 = '';
@array7 = '';
$nohits = 0;      # 0 for print, 1 for don't print (no hits)
@det = '';
$num = 0;
$zfstart = 0;
$zfend = 0;
$ckstart = 0;
$ckend = 0;
$inline = 0;       # 0 or 1
$number = 1;
$IDlength = 0;
$optionCK = 0; #  1 / 2 / 3 / 4 / 5 / 6
$optionZF = 0;
$mid1zf = 0;
$mid2zf = 0;
$mid1ck = 0;
$mid2ck = 0;

open (OUT1, ">./ckzf1380details");
open (OUT3, ">./ckzf1380shortdna.fa");
open (OUT2, ">./ckzf1380protein.fa");
open (OUT4, ">./temp4");
open (OUT5, ">./temp5");
open (OUT6, ">./temp6");
open (OUT7, ">./temp7");
open (IN, "BLASTX.output") || die;
print OUT1 " ZF name\t CK name\tZF DNA\tCK DNA\tE-val1\tScore1\tID1\tFrame1\tE-
val2\tScore2\tID2\tFrame2";

while ($line = <IN>) {
    chomp $line;
    if ($line =~ /No hits found/) { $nohits = 1;}   # remove the 61 with no hits

    if ($line =~ /Query=/) {                         # get ZF name
        @zftemp1 = split /\s+/, $line;
        $zftemp1[1] =~ s/\s+//g;
        $zfname = $zftemp1[1];
        if ($nohits == 0) {
            $count1++;
            print OUT1 "\n$zfname";
            $zfstart = 0;                             # re-set parameters
            $zfend = 0;
            $frame1 = 0;
            $frame2 = 0;
            $inline = 0;
            $number = 0;
            $optionCK = 0;
            $optionZF = 0;
            $mid1zf = 0;
            $mid2zf = 0;
            $IDlength1 = 0;
            $IDlength2 = 0;
            $identity1 = 0;
            $identity2 = 0;        }
        $nohits = 0;                                  # re-set no hits tag
    }

    if (($line =~ />NM_/) || ($line =~ />XM_/)) {   # get CK name
        $line =~ s/\s+//g;
        $line =~ s/>//g;
        $ckname = $line;
        $det[$count1][0] = $zfname;     # NB !!!!!!!!!!!!!!!! ************
        print OUT1 "\t$ckname";
        print OUT1 "\t$zflength";       # needed for DNA seqs
        $det[$count1][1] = $ckname;
        $count2++;
        $ckstart = 0;
        $ckend = 0;
        $mid1ck = 0;
        $mid2ck = 0;
        $strand = 0; # +ve or -ve strand
    }

    if ($line =~ /Score =/) {
        @array1 = split /\s+/, $line;
        $evalu = $array1[8];                        # get e-value
        $score = $array1[3];                        # get score
        print OUT1 "\t$evalu\t$score";
        $det[$count1][2] = $evalu;
        $det[$count1][3] = $score;
    }

    if ($line =~ /Frame =/) {
        @array2 = split /\s+/, $line;
        if ($frame1 == 0) {
            $frame1 = $array2[3];                        # get frame
            print OUT1 "\t$frame1";
            $det[$count1][4] = $frame1;
            if ($frame1 > 0)  { $det[$count1][999] = $frame1 -1  ; }
            if ($frame1 < 0) { $det[$count1][999] = $frame1 +1  ; }
            $inline = 0;
            if ($frame1 < 0) { $strand = -1;
                               $neg++;} # -ve
            else { $strand = 1;
              $pos++;}             # +ve
        }
        elsif ($frame1 != $array2[3]) {
            $frame2 = $array2[3];
            $det[$count1][11] = $frame2;
            print OUT1 "\t$frame2";
            $det[$count1][54] = $frame2;
            if ((($frame1 < 0) && ($frame2 < 0)) || (($frame1 > 0) && ($frame2 > 0))) {
    $det[$count1][999] = $frame2 - $frame1; # +ve, move up 1/2/3 space; -ve move back
                $inline = 1; }
            else { $inline = 0; }                   # ie - if in same frame
        }
        $det[$count1][70] = $inline; # 0 or 1 depending on if multiple hits
        print OUT1 "\tINL=$inline";
    }

    if (($line =~ /letters/) && ($line =~ /\(/)) {
        $line =~ s/\(//g;
        @array3 = split /\s+/, $line;
        $zflength = $array3[1];                 # get length of zf seq
        $det[$count1][5] = $zflength;
    }

    if ($line =~ /Length/) {
        @array4 = split /\s+/, $line;
        $cklength = 3*$array4[3];                   # (NB protein!)
        print OUT1 "\t$cklength";         # get length of ck seq  # needed for DNA seqs
        $det[$count1][6] = $cklength;
    }

    if ($line =~ /Query:/) {
        @array5 = split /\s+/, $line;
        $zfprotpart = $array5[2];                        # get line of ZF protein
        $zfprotpart =~ s/-//g;
        $zfprotpart =~ s/\*//g;

        if (($array5[1] - $zfend) == 1) { $inline = 0; }

        elsif ($inline == 1) {
           if (($array5[1] >= $mid1zf) && ($array5[3] <= $mid2zf)) {   # if in middle of protein
                $midd = ($mid1zf/3 - $zfstart/3 +1);
                $tempZ2             = substr ($det[$count1][10], 0, $midd);
                $det[$count1][10] = substr ($det[$count1][10], $midd, $zfend/3);
                $det[$count1][10] = $tempZ2.$zfprotpart.$det[$count1][10];
                $optionZF = 7;           # special!  !
                $mid3zf = 0;
                $mid4zf = 0;
                if (($array5[1] - $mid1zf) > 1) {
                    $tempZ            = substr ($det[$count1][10], 0, $array5[1]);
                    $det[$count1][10] = substr ($det[$count1][10], $array5[1]/3, $zfend/3);
                    $det[$count1][10] = $tempZ.$det[$count1][10];
                    $mid3zf = $array5[1];                 } # remove extra from protein
                if (($array5[1] - $mid1zf) < 1) {
                    $tempZ            = substr ($det[$count1][10], 0, $array5[1]/3);
                    $det[$count1][10] = substr ($det[$count1][10],(2*$mid1zf/3-
$array5[1]/3),$zfend/3);
```

```
                    $det[$count1][10] = $tempZ.$det[$count1][10];  } # remove overlap from protein
                if (($mid2zf - $array5[3]) > 1) {
                    $tempZ            = substr ($det[$count1][10],0,$array5[3]/3-
$det[$count1][999]/3);
                    $det[$count1][10] = substr ($det[$count1][10], $mid2zf/3, $zfend/3);
                    $det[$count1][10] = $tempZ.$det[$count1][10];
                    $mid4zf = $array5[3];                } # remove extra
                if (($mid2zf - $array5[3]) < 1) {
                    $tempZ            = substr ($det[$count1][10], 0, $array5[3]/3);
                    $det[$count1][10] = substr ($det[$count1][10],(2*$mid2zf/3-
$array5[3]/3),$zfend/3);
                    $det[$count1][10] = $tempZ.$det[$count1][10];                 } # remove overlap
               print OUT6
"\n$zfname\t$ckname\t($zfstart,$mid1zf)\t($mid2zf,$zfend)\t$det[$count1][10]";
             }
           # if(($zfend>=$array5[3])&&($zfstart<=$array5[1])){$optionZF=4;}#option4
            if (($zfstart >=  $array5[1]) && ($zfstart > $array5[3])) {
                $det[$count1][10] = $zfprotpart.$det[$count1][10];
                $optionZF = 6;
                $mid1zf = $array5[3];
                $mid2zf = $zfstart ;
                $zfstart = $array5[1];             } # option 6
            elsif (($zfend < $array5[3]) && ($zfstart >= $array5[1])) {
                $det[$count1][10] = $zfprotpart;
                $optionZF = 5;
                $zfend = $array5[3];
                $zfstart = $array5[1];             } # option 5 (ie replace)
            elsif (($zfend > $array5[3]) && ($zfstart >= $array5[1])) {
                $moddy3 = ($zfstart - $array5[1])%(3);  # change +3 on $end1 HERE !!
                $end1 = ($zfstart - $array5[1] - $moddy3 +3)/3;  # round up to include
                $zfprotpart = substr ($zfprotpart, 0, $end1) ;  # option 3 remove overlap
                $det[$count1][10] = $zfprotpart.$det[$count1][10];
                $optionZF = 3;
                $zfstart = $array5[1];             }
            elsif (($zfend < $array5[1]) && ($zfend < $array5[3])) {
                $det[$count1][10] = $det[$count1][10].$zfprotpart;  # option 1
                $optionZF = 1;
                $mid1zf = $zfend;
                $mid2zf = $array5[1];
                $zfend = $array5[3];            }
            elsif (($zfend >= $array5[1]) && ($zfstart <= $array5[1]) && ($zfend < $array5[3])) {
                $moddy1 = ($zfend - $array5[1])%(3);                  # TOGGLE +1 to 0 / -1
                $divide2 = ($zfend - $array5[1] - $moddy1)/3;
                $moddy2 = ($array5[3] - $array5[1])%(3);             # round down
                $end2 = ($array5[3] - $array5[1] + (3 - $moddy2))/3;  # round up
                $zfprotpart = substr ($zfprotpart, $divide2, $end2);  # option 2 remove overlap
                $det[$count1][10] = $det[$count1][10].$zfprotpart;
                $optionZF = 2;
                $mid1zf = $zfend;
                $mid2zf = $zfend;
                $zfend = $array5[3];            }
        }         # end frame check

        if (($inline == 0) && ($zfstart != 0)) {
            $det[$count1][10] = $det[$count1][10].$zfprotpart;
            $zfend = $array5[3];            }

        if ($zfstart == 0) {
            $zfstart = $array5[1];       # first position
            $zfend = $array5[3];
            $det[$count1][10] = $zfprotpart;        }

        $det[$count1][8]  = $zfstart;
        $det[$count1][9]  = $zfend;
        $det[$count1][30] = $optionZF;
        $det[$count1][31] = $mid1zf - $det[$count1][999] - 1;  # TOGGLE
        $det[$count1][32] = $mid2zf + $det[$count1][999] ;
        $det[$count1][99] = $strand;
        if ($mid3zf != 0 ) {       # remove bit bw 1 & 3
            $det[$count1][33] = $mid3zf - $det[$count1][999] - 1;
            $det[$count1][31] = $mid1zf + $det[$count1][999];        }
        else { $det[$count1][31] = $mid1zf - $det[$count1][999] - 1;
               $det[$count1][32] = $mid2zf + $det[$count1][999] ; }  # TOGGLE
        if ($mid4zf != 0 ) {       # remove bit bw 2 & 4
            $det[$count1][34] = $mid4zf + $det[$count1][999];
            $det[$count1][32] = $mid2zf - $det[$count1][999] - 1;        }
        else { $det[$count1][32] = $mid2zf + $det[$count1][999];
               $det[$count1][31] = $mid1zf - $det[$count1][999] - 1;}
    }

    if ($line =~ /Sbjct:/) {
        @array6 = split /\s+/, $line;                # NB multiply by 3
        $ckprotpart = $array6[2];                        # get line of CK protein
        $ckprotpart =~ s/-//g;
        $ckprotpart =~ s/\*//g;

        if ($array6[1] - $ckend == 1) { $inline = 0; }
        elsif ($inline == 1) {
            if (($array6[1] >= $mid1ck/3) && ($array6[3] <= $mid2ck/3)) {
                $mid = ($mid1ck/3 - $ckstart +1 );
                $tempCS             = substr ($det[$count1][20], 0, $mid);
                $det[$count1][20] = substr ($det[$count1][20], $mid, $ckend);
                $det[$count1][20] =$tempCS.$ckprotpart.$det[$count1][20];# +1 cos of substr method
                $optionCK = 7;          # special!  !
                $mid3ck = 0;
                $mid4ck = 0;
                if (($array6[1] - $mid1ck/3) > 1) {
                    $tempC          = substr ($det[$count1][20], 0, $array6[1]);
                    $det[$count1][20] = substr ($det[$count1][20], $array6[1], $ckend);
                    $det[$count1][20] = $tempC.$det[$count1][20];   } # remove extra from protein
                if (($array6[1] - $mid1ck/3) < 1) {
                    $tempC          = substr ($det[$count1][20], 0, $array6[1]);
                    $det[$count1][20] = substr ($det[$count1][20], (2*$mid1ck/3 - $array6[1]),
$ckend);
                    $det[$count1][20] = $tempC.$det[$count1][20];
                    $mid3ck = 3*$array6[1];                } # remove overlap from protein
                if (($mid2ck/3 - $array6[3]) > 1) {
                    $tempC          = substr ($det[$count1][20], 0, $array6[3]);
                    $det[$count1][20] = substr ($det[$count1][20], $array6[3], $ckend);
                    $det[$count1][20] = $tempC.$det[$count1][20];                } # remove extra
                if (($mid2ck/3 - $array6[3]) < 1) {
                    $tempC          = substr ($det[$count1][20], 0, $array6[3]);
                    $det[$count1][20] = substr ($det[$count1][20], (2*$mid2ck/3 - $array6[3]),
$ckend);
                    $det[$count1][20] = $tempC.$det[$count1][20];
                    $mid4ck = 3*$array6[3];                } # remove overlap
                print OUT6 "\n$ckname\t$zfname\t($ckstart,",$mid1ck/3,")\t(",$mid2ck/3,",$ckend)";
                print OUT6 "\t$array6[1],$array6[3]\n$tempC\n$det[$count1][20]";
            }
            elsif (($ckstart > $array6[3]) && ($ckstart >= $array6[1])) {
                $det[$count1][20] = $ckprotpart.$det[$count1][20];
                $optionCK = 6;
                $mid1ck = 3*$array6[3];
                $mid2ck = 3*$ckstart;
                $ckstart = $array6[1];              } # option 6
            if (($ckend < $array6[3])&& ($ckstart >=  $array6[1])) {
                $det[$count1][20] = $ckprotpart;
                $ckend = $array6[3];
                $optionCK = 5;
                $ckstart = $array6[1];               } # option 5 (ie replace)
            elsif (($ckend > $array6[3]) && ($ckstart >=  $array6[1])) {
                $end11 = ($ckstart - $array6[1]);           # no need to divide by 3
                $ckprotpart = substr ($ckprotpart, 0, $end11) ; # option 3 remove overlap
                $det[$count1][20] = $zfprotpart.$det[$count1][20];
                $optionCK = 3;
                $ckstart = $array6[1];            }
            elsif (($ckend < $array6[1]) && ($ckend < $array6[3])) {
                $det[$count1][20] = $det[$count1][20].$ckprotpart;
                $optionCK = 1;
                $mid1ck = 3*$ckend;
                $mid2ck = 3*$array6[1];
                $ckend = $array6[3];             } # option 1
            elsif (($ckend >= $array6[1]) && ($ckstart <= $array6[1])&& ($ckend < $array6[3])) {
                $divide22 = $ckend - $array6[1];      # TOGGLE +1 to 0 / -1
                $end22 = $array6[3] - $array6[1];
                $ckprotpart = substr ($ckprotpart, $divide22, $end22);   # option 2 remove middle
                $det[$count1][20] = $det[$count1][20].$ckprotpart;
                $optionCK = 2;
                $mid1ck = 3*$ckend;
                $mid2ck = 3*$ckend;
                $ckend = $array6[3];           }
        }           # end frame check

        if (($inline == 0) && ($ckstart != 0)) {
            $det[$count1][20] = $det[$count1][20].$ckprotpart;
            $ckend = $array6[3];
        }
        if ($ckstart == 0) {
            $ckstart = $array6[1];
            $ckend = $array6[3];
            $det[$count1][20] = $ckprotpart;  }     # first position
        $det[$count1][18] = $ckstart;
        $det[$count1][19] = $ckend;
        $det[$count1][40] = $optionCK;
        $det[$count1][100] = $strand;
        $det[$count1][41] = $mid1ck - $det[$count1][999] - 1;  # TOGGLE
        $det[$count1][42] = $mid2ck + $det[$count1][999] ;
        if ($mid3ck != 0 ) {
            $det[$count1][43] = $mid3ck - $det[$count1][999] - 1; # gap bw 1 & 3
            $det[$count1][41] = $mid1ck + $det[$count1][999];       }
        else { $det[$count1][41] = $mid1ck - $det[$count1][999] - 1;
                $det[$count1][42] = $mid2ck + $det[$count1][999] ; }  # TOGGLE
        if ($mid4ck != 0 ) {
            $det[$count1][44] = $mid4ck + $det[$count1][999];
            $det[$count1][42] = $mid2ck - $det[$count1][999] - 1;       }
        else { $det[$count1][42] = $mid2ck + $det[$count1][999];
               $det[$count1][41] = $mid1ck - $det[$count1][999] - 1;}   # TOGGLE
    }

    if ($line =~ /Identities/) {
        $line =~ s/\(//g;
        $line =~ s/\)//g;
        $line =~ s/\,//g;
        $line =~ s/\%//g;
        @array7 = split /\s+/, $line;
        if ($identity1 == 0) {
            $identity1 = $array7[4];
            $det[$count1][7] = $identity1;
            if ($identity1 >=70) { print OUT1 "\t$identity1%";}
            else { print OUT1 "\t$identity1"; }
            $det[$count1][56] = 100;         # for check later
            $inline = 0; }                   # get identity (as a %)
        else { $inline = 1;
               $identity2 = $array7[4];
               $det[$count1][56] = $identity2; }

        @array8 = split /\//, $array7[3];
        if ($IDlength1 == 0) {
            $inline = 0;
            $IDlength1 = $array8[1];
            if ($IDlength1 >= 65) { print OUT1 "\tL=$IDlength1";}
            else { print OUT1 "\tl=$IDlength1"; }
            $det[$count1][13] = $IDlength1;
            $det[$count1][55] = 100;    # for check-point
        }
        else {$inline = 1;
              $IDlength2 = $array8[1];
              $det[$count1][55] = $IDlength2;
              if ($IDlength2 >= 65) { print OUT1 "\tL=$IDlength2";}
              else { print OUT1 "\tl=$IDlength2"; }
        }     # second one
    }
}

########################## get DNA seqs & put into file separately ##################################

for ($i=1; $i < $count1+1; $i++) {           # goes from 1 - 1380 too; 0th is empty
  $stw="select ckALLseqs.ckname, ckALLseqs.ckseq from ckALLseqs where ckALLseqs.ckname =
'$det[$i][1]';";
    $sth = $dbh->prepare($stw) or die "Can't prepare $stw: $dbh->errstr\n";
    $rv = $sth->execute() or die "Can't execute the query: $sth->errstr";

    if ($rv eq 1)   {
        while(@array=$sth->fetchrow_array)  {
            $chickenname = $array[0];
            $chickenseq = $array[1];
            $Cdna[0][$i] = $chickenname;
            $Cdna[1][$i] = $chickenseq;          }
    }

    $stw="select zf5661DNAcontigs.zfname, zf5661DNAcontigs.zfseq from zf5661DNAcontigs where
zf5661DNAcontigs.zfname = '$det[$i][0]';";
    $sth = $dbh->prepare($stw) or die "Can't prepare $stw: $dbh->errstr\n";
    $rv = $sth->execute() or die "Can't execute the query: $sth->errstr";

    if ($rv eq 1)   {
        while(@array=$sth->fetchrow_array)  {
            $zebname = $array[0];
            $zebseq = $array[1];
            $Cdna[2][$i] = $zebname;
            $Cdna[3][$i] = $zebseq;     }
    }
    $checkme1 = 0;
    $checkme2 = 0;

    if (($det[$i][99] < 0) && ($det[$i][8] > $det[$i][9])) {        # check if strand is -ve
        $checkme1 = 1;
        $tempk = $det[$i][9];
        $det[$i][9] = $det[$i][8];                # zf
        $det[$i][8] = $tempk;
        $tempk = $det[$i][32];
        $det[$i][32] = $det[$i][31];
        $det[$i][31] = $tempk;    }

    if (($det[$i][100] < 0) && ($det[$i][18] > $det[$i][19])) {        # check if strand is -ve
        $checkme2 = 1;
        $tempk = $det[$i][19];
        $det[$i][19] = $det[$i][18];              # ck
        $det[$i][18] = $tempk;
        $tempk = $det[$i][42];
        $det[$i][42] = $det[$i][41];
        $det[$i][41] = $tempk;    }
    if ($det[$i][18] < 1) { $det[$i][18] = 0;}    # adjust for arrays starting from 0 not 1
    else { $det[$i][18] = 3*$det[$i][18] - 3; }
    $det[$i][19] = 3*$det[$i][19];
    if ($det[$i][8] < 1) { $det[$i][8] = 0;}
    else { $det[$i][8]--; }
    $lengthZF = (length $Cdna[3][$i]);
    $lengthCK = (length $Cdna[1][$i]);

    if ($checkme1 == 1) { $Cdna[3][$i] = reverse ($Cdna[3][$i]);       # if -ve reverse
                          $Cdna[3][$i] =~ tr/ACGT/TGCA/; }          # and swop
    if ($checkme2 == 1) { $Cdna[1][$i] = reverse ($Cdna[1][$i]);
                           $Cdna[1][$i] =~ tr/ACGT/TGCA/;  }

    if (($det[$i][30] == 1)||($det[$i][30]==2)||($det[$i][30]==6)) { # if options 1 or 6
        $tempZF      = substr ($Cdna[3][$i], $det[$i][32],  ($det[$i][9]-$det[$i][32]));
        $Cdna[3][$i] = substr ($Cdna[3][$i], $det[$i][8], ($det[$i][31]-$det[$i][8]));
        $Cdna[3][$i] = $Cdna[3][$i].$tempZF;
    }
    elsif ($det[$i][30] == 7) {
        if ($det[$i][33] != 0) {     # bw 1 and 3
            $tempZF      = substr ($Cdna[3][$i], $det[$i][33],  ($det[$i][9]-$det[$i][33]));
            $Cdna[3][$i] = substr ($Cdna[3][$i], $det[$i][8], ($det[$i][31]-$det[$i][8]));
            $Cdna[3][$i] = $Cdna[3][$i].$tempZF;         }
        if ($det[$i][34] != 0) {     # bw 2 and 4
            $tempZF      = substr ($Cdna[3][$i], $det[$i][34],  ($det[$i][9]-$det[$i][34]));
            $Cdna[3][$i] = substr ($Cdna[3][$i], $det[$i][8], ($det[$i][32]-$det[$i][8]));
            $Cdna[3][$i] = $Cdna[3][$i].$tempZF;         }
    }
    else { $Cdna[3][$i] = substr ($Cdna[3][$i], $det[$i][8], $det[$i][9]);}

    if (($det[$i][40]==1)||($det[$i][40]==6)||($det[$i][40]==2)){# same for CK - remove "middle" of
seq
        $tempCK      = substr ($Cdna[1][$i], $det[$i][42], $det[$i][19]);
        $Cdna[1][$i] = substr ($Cdna[1][$i], $det[$i][18], $det[$i][41]);
        $Cdna[1][$i] = $Cdna[1][$i].$tempCK;    }
    elsif ($det[$i][40] == 7) {
        if ($det[$i][43] != 0) {     # bw 1 and 3
            $tempCK      = substr ($Cdna[1][$i], $det[$i][43], ($det[$i][19]-$det[$i][43]));
            $Cdna[1][$i] = substr ($Cdna[1][$i], $det[$i][18], ($det[$i][41]-$det[$i][18]));
            $Cdna[1][$i] = $Cdna[1][$i].' | '.$tempCK;         }
        if ($det[$i][44] != 0) {     # bw 2 and 4
            $tempCK      = substr ($Cdna[1][$i], $det[$i][44], ($det[$i][19]-$det[$i][44]));
            $Cdna[1][$i] = substr ($Cdna[1][$i], $det[$i][18], ($det[$i][42]-$det[$i][18]));
            $Cdna[1][$i] = $Cdna[1][$i].' Â£ '.$tempCK;         }
    }
    else {$Cdna[1][$i]= substr ($Cdna[1][$i], $det[$i][18], $det[$i][19]);}  # CK

    if (($det[$i][13] >= 65) && ($det[$i][55] >= 65)) {
        if (($det[$i][7] >= 70) && ($det[$i][56] >= 70)) {
            if (($Cdna[2][$i] =~ $det[$i][0]) && ($Cdna[0][$i] =~ $det[$i][1])) {
                print OUT2 "\nÂ£>$det[$i][0]\n$det[$i][10]";                                  #  protein
names
                print OUT2 "\n>$det[$i][1]\n$det[$i][20]";             # protein seqs
                print OUT3 "\nÂ£>$Cdna[2][$i]\n$Cdna[3][$i]";             # zf + ck DNA names
                print OUT3 "\n>$Cdna[0][$i]\n$Cdna[1][$i]";
                $det[$i][10] =~ s/x//g;
                $det[$i][20] =~ s/x//g;
                $det[$i][10] =~ s/B//g;
                $det[$i][20] =~ s/B//g;
                $det[$i][10] =~ s/O//g;
                $det[$i][20] =~ s/O//g;
                $det[$i][10] =~ s/U//g;
                $det[$i][20] =~ s/U//g;
                $det[$i][10] =~ s/J//g;
                $det[$i][20] =~ s/J//g;
                open (OUTP, ">./prot$i.fa");                        # 1k protein seqs
                print OUTP ">$det[$i][0]\n$det[$i][10]\n>$det[$i][1]\n$det[$i][20]\n";
                open (OUTD, ">./dna$i.fa");                         # 1k dna seqs
                print OUTD ">$Cdna[2][$i]\n$Cdna[3][$i]\n>$Cdna[0][$i]\n$Cdna[1][$i]\n";

                system ("transeq dna$i.fa transeq$i.fa -auto"); # TRANSEQ; names have "_1" attached
                open (IN7, "transeq$i.fa") || die;
                @seq = <IN7>;
                $input7 = join ('', @seq);
                $input7 =~ s/\s+//g;
                $input7 =~ s/>/\n/g;
                $input7 =~ s/_1/_1\n/g;
                $input7 =~ s/_1//g; # removes from zf only
                $input7 =~ s/\*//g;
                $input7 =~ s/x//g;
                $input7 =~ s/B//g;
                $input7 =~ s/O//g;
                $input7 =~ s/U//g;
                $input7 =~ s/J//g;
                @seq = split /\n/, $input7;   # zf protein = $seq[2], ck protein = $seq[4]
                if ($seq[2] =~ $det[$i][10]) { if ($det[$i][30]==0) { $z1f0++; }
                                               if ($det[$i][30]==1) { $z1f1++; }
                                               if ($det[$i][30]==2) { $z1f2++; }
                                               if ($det[$i][30]==3) { $z1f3++; }
                                               if ($det[$i][30]==4) { $z1f4++; }
                                               if ($det[$i][30]==5) { $z1f5++; }
                                               if ($det[$i][30]==6) { $z1f6++; }
                                               $ernie++;
                                              # print OUT4 "\n$i\t$seq[1]";   # check ZF name
                                           }   # print OUT4 "\nT=$seq[2]"; }  # check ZF seq
                else { if ($det[$i][30]==0) { $z0f0++; }
                       if ($det[$i][30]==1) { $z0f1++; }
                       if ($det[$i][30]==2) { $z0f2++; }
                       if ($det[$i][30]==3) { $z0f3++; }
                       if ($det[$i][30]==4) { $z0f4++; }
                       if ($det[$i][30]==5) { $z0f5++; }
                       if ($det[$i][30]==6) { $z0f6++; }
                       print OUT5 "\n$seq[1]\toption=$det[$i][30]\tstrand=$det[$i][99]";
                       print OUT5 "\tFrameDiff=$det[$i][999]\tzfst=$det[$i][8]\tend=$det[$i][9]";
                       print OUT5 "\tm1=$det[$i][31]\tm2=$det[$i][32]\t=$det[$i][30]\tckm=$checkme1";
                       print OUT5 "\nS=$seq[2]\nR=$det[$i][10]";
                       print OUT5 "\nSD=$Cdna[3][$i]\n________________";}    # zf
                if ($seq[4] =~ $det[$i][20]) { $bert++;
                                               if ($det[$i][40]==0) { $c1k0++; }
                                               if ($det[$i][40]==1) { $c1k1++; }
                                               if ($det[$i][40]==2) { $c1k2++; }
                                               if ($det[$i][40]==3) { $c1k3++; }
                                               if ($det[$i][40]==4) { $c1k4++; }
                                               if ($det[$i][40]==5) { $c1k5++; }
                                               if ($det[$i][40]==6) { $c1k6++; }
                                           }   # print OUT4 "\n$i\t$seq[3]";   # Ck name
                                              # print OUT4 "\nT=$seq[4]"; }  # Ck seq
                else { if ($det[$i][40]==0) { $c0k0++; }
                       if ($det[$i][40]==1) { $c0k1++; }
                       if ($det[$i][40]==2) { $c0k2++; }
                       if ($det[$i][40]==3) { $c0k3++; }
                       if ($det[$i][40]==4) { $c0k4++; }
                       if ($det[$i][40]==5) { $c0k5++; }
                       if ($det[$i][40]==6) { $c0k6++; }
                       print OUT5 "\n$seq[3]\t$seq[1]\toption=$det[$i][40]\tstrand=$det[$i][100]";
                       print OUT5 "\tFrameDiff=$det[$i][999]\tckst=$det[$i][18]\tend=$det[$i][19]";
                       print OUT5 "\tm1=$det[$i][41]\tm2=$det[$i][42]\t=$det[$i][40]\tckm=$checkme2";
                       print OUT5 "\nS=$seq[4]\nR=$det[$i][20]";
                       print OUT5 "\nSD=$Cdna[1][$i]\n___________________"; }       # ck

                #system ("t_coffee -infile=prot$i.fa -outfile tcpt$i.aln"); # T-COFFEE
            }
        }
    }
}
print "\nNos zf = $ernie\tck = $bert";
print "\nPos = $pos\tNeg = $neg\tRe-setzf=$tttt\tck=$qqqq";
print "\n       \t0\t1\t2\t3\t4\t5\t6";
print "\nDodgeCK:\t$c0k0\t$c0k1\t$c0k2\t$c0k3\t$c0k4\t$c0k5\t$c0k6";
print "\nDodgeZF:\t$z0f0\t$z0f1\t$z0f2\t$z0f3\t$z0f4\t$z0f5\t$z0f6\n";

###################### separate - DNA file ##################

open (IN3,"ckDNAall.fa") || die;       # contains 24 k seqs
@a = <IN3>;                                 # DNA seqs
$input3 = join ('', @a);
$input3 =~ s/\_cds\_1//g;
$input3 =~ s/\_cds\_2//g;
$input3 =~ s/\_cds\_3//g;
$input3 =~ s/\_cds\_4//g;
$input3 =~ tr/actgnxm/ACTGNXM/;
$input3 =~ s/>/\nÂ£>/g;
@a = split /Â£/, $input3;

for ($i=1; $i < scalar @a; $i++) {  # 0th element is empty
    $a[$i] =~ s/\n/|/o;
    $a[$i] =~ s/\n//g;
    $a[$i] =~ s/>/\n/g;
    $a[$i] =~ s/\|/\t/g;    }           # make file MySQL - readable

open (IN4,"zebraf.contigs") || die;   # has 5661 contigs
@b = <IN4>;                                 # DNA seqs
$input4 = join ('', @b);
$input4 =~ tr/aclxm/ACLXM/;
$input4 =~ s/>/\nÂ£>/g;
@b = split /Â£/, $input4;

for ($i=1; $i < scalar @b; $i++) {  # 0th element is empty
    $b[$i] =~ s/\n/|/o;
    $b[$i] =~ s/\n//g;
    $b[$i] =~ s/>/\n/g;
    $b[$i] =~ s/\|/\t/g;   }            # make file MySQL - readable
                                       #... Then manually load data into MySQL ...
exit;
```

hitParserZF.pl

```
#!/usr/bin/perl
# Program to parse Blastx output from turkey/zebra finch - chicken blastx, getting:
# name (ck refseq), e-value, score, length (ck), description (ck), #HSPs,
# frame (for each HSP), contig name, contig length.
# Then only keep seqs with 71+ aAs and % identity > 60%
# Then get positions: ck = sbjct = aA; tk/zf = query = DNA (so x3) - check frame
# and check orientation - may have to merge HSPs later.
# Make export data for MySQL for names, etc.

use DBI;
$DBD      = 'mysql';
$host     = 'popgen.gen.tcd.ie';
$user     = 'downingt';
$password = 'peeso5ni';
$database = 'tim';
$dbh = DBI->connect("DBI:$DBD:$database:$host","$user","$password", { RaiseError => 1, AutoCommit => 1});

open (BLASTX, "ZFcontigsJanvCKprotRSBLASTX.out") || die "Cannot open blastxfile\n";
open (IN, "zebfin-all.fa.contigs") || die "Can't open tk\n";
open (OUT2, ">./temp1");
open (OUT1, ">./blastxZFOutput.txt") || die "Cannot open templfile\n";
@blastx = <BLASTX>;
$input1 = join ('',@blastx);
@blastx = split /Reference/, $input1;  # element 0 = nothing
@contigs = <IN>;
$input2 = join ('',@contigs);
@contigs = split />/, $input2; # from 1...1811

# Remove ones with no hits, remove ones with #aAs < 70 & identities < 60%, put good ones in array
$count1 = 0;
$count2 = 0;
$n = 1;
for ($i=1; $i < scalar @blastx; $i++) {
    if ($blastx[$i] =~ "No hits found") { $count1++; }
    else {
        @hsps = split /Score =/, $blastx[$i];
        # [0] = names, eval, etc [1] = hsp 1 [2] = hsp 2 (if present) ...
        @array1 = split /\n/, $hsps[1];  # get 1st HSP, split each line
        if ($array1[1] =~ /Identities =/) {
            @ident = split /\s+/, $array1[1];
            $ident[4] =~ s/\(//g;
            $ident[4] =~ s/\)//g;
            $ident[4] =~ s/\%//g;
            $ident[4] =~ s/\,//g;   # check if identities % > 60%
            if ($ident[4] > 60) {
                @long = split /\s+/, $array1[4];
                @long2 = split /\s+/, $array1[6];
                $sum = length($long[2]) + length($long2[2]);
                if ($sum > 70) { # if length of 1st HSP > 70 aAs
                    $blast[$n] = $blastx[$i];     # put into array for analysis
                    $n++;
                    $count2++; }            }        }   }
}

print "\n# with no hits = $count1\n#okay = $count2\n";
$x=1;
for ($i=1; $i < scalar @blast; $i++) {
    @hsps = split /Score =/, $blast[$i]; # [0] = names, eval, etc [1] = hsp 1 [2] = hsp 2 (if
present)..
    @array = split /\n/, $hsps[0];
    for ($g=0; $g < scalar @array; $g++) {
        if ($array[$g] =~ /Query=/) {
            $array[$g+1] =~ s/\(//g;
            $array[$g+1] =~ s/\)//g;
            $array[$g+1] =~ s/[lettrs]//g;
            $array[$g+1] =~ s/\s+//g; # query name + length
            $array[$g] =~ s/[Query= ]//g;
            $queryName = $array[$g];
            print OUT1 "\n>Query Name = $array[$g]\tlength contig = $array[$g+1]\t#$i\n"; }
        elsif ($array[$g] =~ />/) { $array[$g] =~ s/>//g;
                                    $array[$g+1] =~ s/\s+//g;  # refseq name + description
                                    @ainm = split /\|/, $array[$g];
                                    @nom = split /\./, $ainm[1];
                                    $subjectName = $nom[0];
                                    $stuff[$i] = $nom[0];
                                 #   if (!($nom[0] =~ "P_")) { print
                                    if($array[$g+1]=~/Length/) { print OUT1 "n=$nom[0]\t$array[$g]\n"; }
                                    else { print OUT1 "n=$nom[0]\t$array[$g]$array[$g+1]\n"; }
                                }
        elsif ($array[$g] =~ /Length/) { $array[$g] =~ s/[Length=]//g;
                                         $array[$g] =~ s/\s+//g; # refseq length
                                           print OUT1 "length refseq = $array[$g]\n";  }    }
     for ($r=1; $r < scalar @hsps; $r++) {
        $CKseq = '';
        $TZseq = '';
        $TZstartcheck = 0;
        $CKstartcheck = 0;
        @gethsp = split /\n/, $hsps[$r];
        for ($g=0; $g < scalar @gethsp; $g++) {
            if ($gethsp[$g] =~ /Expect/) { @scorey = split /\s+/, $gethsp[$g];
                                           print OUT1 "Score = $scorey[1]\tE-value = $scorey[6]"; }
            if ($gethsp[$g] =~ /Identities/) { @identity = split /\s+/, $gethsp[$g];
                                               $identity[4] =~ s/\)//g;
                                               $identity[4] =~ s/\(//g;
                                               $identity[4] =~ s/\%//g;
                                               $identity[4] =~ s/\,//g;
                                               print OUT1 "\tIdentities = $identity[4]%"; }
            if ($gethsp[$g] =~ /Frame/) { @frame = split /\s+/, $gethsp[$g];
                                          $frm = $frame[3];
                                          print OUT1 "\tframe = $frame[3]"; }
            if ($gethsp[$g] =~ /Query:/) { @query = split /\s+/, $gethsp[$g];
                                           $TZseq = $TZseq.$query[2];
                                           if ($TZstartcheck == 0) { $startTZ = $query[1]; }
                                           $TZstartcheck = 1;
                                           $endTZ = $query[3]; }
            if ($gethsp[$g] =~ /Sbjct:/) { @subject = split /\s+/, $gethsp[$g];
                                           $CKseq = $CKseq.$subject[2];
                                           if ($CKstartcheck == 0) { $startCK = $subject[1]; }
                                           $CKstartcheck = 1;
                                           $endCK = $subject[3];    }        }

        # look up protein name in MySQL to get refseq one
        $stw="select chickenRefSeqPositions.name from chickenRefSeqPositions where
chickenRefSeqPositions.proteinname = '$stuff[$i]';";
        $sth = $dbh->prepare($stw) or die "Can't prepare $stw: $dbh->errstr\n";
        $rv = $sth->execute() or die "Can't execute the query: $sth->errstr";
        if ($rv eq 1) { while(@array=$sth->fetchrow_array)  { $correct = $array[0]; }   }
        print OUT1"\n#$correct\tst=$stuff[$i]\t$TZseq\nTZ=$startTZ $endTZ\n$CKseq\nCK=$startCK
$endCK\n";
        open (TEMP, ">./prot_files/prot$x.fa") || die "Cannot open prot $x file\n";
        print TEMP "\n>$queryName\n\n$TZseq\n>$correct\n\n$CKseq";

        if ($startTZ > $endTZ) { # check if -ve frame - no -ve frames in CK
            $temp = $startTZ;
            $startTZ = $endTZ;
            $endTZ = $temp; }  # NB: frame doesn't matter for getting DNA seqs

        for ($y=1; $y < scalar @contigs; $y ++) {
            $tempSeq = '';
            @contig = split /\n/, $contigs[$y];     # does it line by line
            if ($queryName eq $contig[0]) {
                for ($d=1; $d < scalar @contig; $d++) { $tempSeq = $tempSeq.$contig[$d]; }
                if ($frm < 0) {                   # cut then reverse then transpose if -ve frame
                    $tempSeq = substr ($tempSeq, ($startTZ-1), ($endTZ-$startTZ+1));
                    @tempA = split //, $tempSeq;
                    $tempSeq = '';
                    for ($a=scalar @tempA; $a > -1; $a--) { $tempSeq = $tempSeq.$tempA[$a]; }
                    $tempSeq =~ tr/ACGT/TGCA/;
                    $TZcds = $tempSeq;                }
                elsif ($frm > 0) { $TZcds = substr ($tempSeq, ($startTZ-1), ($endTZ-$startTZ+1)); }
                # cross-reference CK protein names to MySQL db to get CDS seq

                $stw="select chickenRefSeqPositions.name, chickenRefSeqPositions.start,
chickenRefSeqPositions.end, chickenRefSeqPositions.codonstart, chickenRe
fSeqPositions.proteinname, chickenRefSeqPositions.chickenCDS from chickenRefSeqPositions where
chickenRefSeqPositions.proteinname = '$stuff[$i]';";
                $sth = $dbh->prepare($stw) or die "Can't prepare $stw: $dbh->errstr\n";
                $rv = $sth->execute() or die "Can't execute the query: $sth->errstr";

                if ($rv eq 1)   {
                    while(@array=$sth->fetchrow_array)  { $ainm = $array[0];
                                                          $st = $array[1];
                                                          $nd = $array[2];
                                                          $codst = $array[3];
                                                          $proteinAinm = $array[4];
                                                          $chickCDS = $array[5];   }                }
                $chickCDS = substr ($chickCDS, ($st-1), ($nd-$st+1));
                $chickCDS = substr ($chickCDS, 3*($startCK-1)+$codst-1, 3*($endCK-$startCK+1)+$codst-1);
                if ((length $chickCDS < 20) || (!($proteinAinm eq $stuff[$i]))) {
                    print OUT2 ">$queryName\n>$stuff[$i]\t$ainm\t",3*($startCK-1),"\t",3*($endCK-
$startCK+1),"\t$proteinAinm\tx = $x\t$codst\n$chickCDS\n$CKseq\n";
                    $chickCDS = "Undefined";                 }
                open (CDS, ">./cds_files/cds$x.fa") || die "Cannot open cds $x file\n";
                print CDS ">$queryName\n$TZcds\n>$ainm\n$chickCDS\n";           }
        } # end for each contig
           $x++;
        for ($m=1;$m<41;$m++){if($x == $m*100){ print "\n",sprintf("%.0f",($m*100)/40),"% done"; } }
    } # for each HSP - so use $x instead of $i
}
print "\nNumber of files = $x\n"; exit;
```

Chapter 4

Perl scripts used in parsing and analysis of chicken EST SNPs: pnpsdets.pl and genbankdets.pl.

pnpsdets.pl

```
#!/var/usr/perl/ # Program to obtain ratios of SNPs and substitution types
use DBI;
$DBD      = 'mysql';
$database = 'test';
$dbh = DBI->connect("DBI:$DBD:$database", { RaiseError => 1, AutoCommit => 1});

open (OUT1, ">./SNPratios.out");
open (OUT2, ">./SNPratios.mysql");
open (OUT3, ">./temp1");
open (IN1, "refseqnames.list") || die;  # File with GenBank data
@codonpos = '';
$snptype = '';
$subtype = '';
@entry = '';

@list = <IN1>;
$input1 = join ('', @list);
@list = split /\n/, $input1;            # divide into each hit
print "\nTotal number of entries = ", scalar @list -1, "\n"; # element 0 is empty
print OUT1 "\nName\t\tPn/Ps\t\tNon-con/Con\tNumber of alleles\tCodon pos";

for($i=1; $i < scalar @list; $i++) { # Need to check through refseq names & add up
   $ns = 0;
   $s = 0;
   $nc = 0;
   $c = 0;
   $syn = 0;
   @entry = split /\s+/, $list[$i];     # name = [0]
   $stw="select SNPhits.refseqname, SNPhits.SNPtype, SNPhits.substitntype, SNPhits.description,
SNPhits.SNPcodonpos from SNPhits where SNPhits.refseqname = '$entry[0]';";
   $sth = $dbh->prepare($stw) or die "Can't prepare $stw: $dbh->errstr\n";
   $rv = $sth->execute() or die "Can't execute the query: $sth->errstr";

  # if ($rv eq 1)   {
   while(@array=$sth->fetchrow_array)  {
       $name = $array[0];               # assign MySQL printout to new array
       $snptype = $array[1];
       $subtype = $array[2];
       $description = $array[3];
       $codonpos[$i] = $array[4];
#       print OUT3 "\n$array[0]\t$array[1]\t$array[2]\t$array[3]\t$array[4]";
       if ($snptype =~ /n/) { $ns++; }
       elsif ($snptype eq "sSNP") { $s++; }
       if ($subtype =~ /C/) { $c++; }
       if ($subtype =~ /c/) { $nc++; }
       elsif ($subtype =~ /S/) { $syn++; }   # syn if neither Con or non-con
   }

   if (($ns != 0) && ($s != 0)) { $nsratio = $ns/$s;
                                  $nsratio = sprintf ("%.3f", $ns/$s);  }
   elsif (($ns != 0) && ($s == 0)) { $nsratio = 999; }
   else { $nsratio = 0; }
   if (($nc != 0) && ($c != 0)) { $ncratio = $nc/$c;
                                  $ncratio = sprintf ("%.3f", $nc/$c); }
   elsif (($nc != 0) && ($c == 0)) { $ncratio = 999; }
   else { $ncratio = 0; }
   $num = $ns + $s +1;                     # number of alleles present

   if (!($name eq /\s+/)) {
    print OUT1 "\n$name\t$ns/$s = $nsratio\t$nc/$c = $ncratio\t$num\t\t$codonpos";
    print OUT2 "\n$name\t$ns\t$s\t$nsratio\t$nc\t$c\t$ncratio\t$num\t$description";} }
exit;
```

genbankdets.pl

```
#!/var/usr/perl/
# Program to obtain details from GenBank info

use DBI;
$DBD      = 'mysql';
$database = 'test';
$dbh = DBI->connect("DBI:$DBD:$database", { RaiseError => 1, AutoCommit => 1});
# Split genbank file into arrrays
@details = '';
@dets = '';
```

```
$j = 0;   # iterates for @dets

#open (OUT1, ">./temp1")|| die;
open (OUT2, ">./snphits.gb")|| die;
open (OUT3, ">./names.list")|| die;
open (IN1, "chicken-rs-mrna.gb") || die;  # File with GenBank data
@genbank = <IN1>;
$input1 = join ('', @genbank);
@genbank = split /LOCUS/, $input1;            # divide into each hit
print "\nTotal number of entries = ", scalar @genbank, "\n"; # [0] element is empty

$version = 100;
#$version = scalar @genbank;
print "Number of used entries = $version";

# Get names of genbank files and put them in a file for mysql

for ($i=1; $i < $version; $i++) {
   $chrom = 999;
   $defin = 999;  # re-set to get details for each entry
   $accessn = 0;
   @getname = split /\s+/, $genbank[$i];   # name = [1]
   $getname[1] =~ s/\s+//g;

   # Get GenBank details: [1] = name [2] = length (bp), then name, chromosome.
 #   print OUT1 "\n$getname[1]\t$getname[2]\t";
   for ($m=1; $m < scalar @getname; $m++) {  if ($getname[$m] =~ /DEFINITION/) { $defin = $m + 1; }
         if ($getname[$m] =~/ACCESSION/) { $accessn = $m; }    }
   for ($zzz = $defin; $zzz < $accessn; $zzz++) { }
#if (!(($getname[$zzz] =~ /allus/)||($getname[$zzz] =~ /RNA/))){ print
OUT1 "$getname[$zzz] "; }

   for ($m=1; $m < scalar @getname; $m++) {
       if ($getname[$m] =~ /chromosome=/) {$chrom = $m;   # use $getname[$chrom];
                                           $getname[$chrom] =~ s/\///g;
                                           $getname[$chrom] =~ s/\"//g; }
                                            #print OUT1 "\t$getname[$chrom]";
       if ($getname[$m] =~ /gene=/) { $genename = $m;
                                      $getname[$genename] =~ s/\"//g;
                                      $getname[$genename] =~ s/gene=//g;
                                      $getname[$genename] =~ s/\///g;   }     }

   # Acess MySQL and compare genbank names to names of refseqs with SNPs in "allSNPs" table

   $stw="select allSNPs.refseqname,  allSNPs.SNPcodonpos, allSNPs.SNPtype, allSNPs.majoraA,
allSNPs.majornum,  allSNPs.majorfreq,    allSNPs.minoraA,       allSNPs.minornum, allSNPs.minorfreq,
  allSNPs.minoreraA, allSNPs.minorernum,    allSNPs.minorerfreq, allSNPs.minorereraA,
allSNPs.minorerernum,allSNPs.minorererfreq, allSNPs.substitntype from allSNPs where
allSNPs.refseqname = '$getname[1]';";
   $sth = $dbh->prepare($stw) or die "Can't prepare $stw: $dbh->errstr\n";
   $rv = $sth->execute() or die "Can't execute the query: $sth->errstr";

   while(@array=$sth->fetchrow_array)  {
       for ($x=0; $x < 16; $x++) { $details[$i][$x] = $array[$x]; } #assign MySQL printout to new
array

       if (!($details[$i][0] eq /\s+/)) {            $j++;            # starts at element [1], like $i
           for ($y=0; $y < 16; $y++) { $dets[$j][$y] = $details[$i][$y]; }

           print OUT2 "\n$dets[$j][0]";     # need to add "\n" at new entry and "\t" for details
           for ($w=1; $w < 16; $w++) { print OUT2 "\t$dets[$j][$w]"; }
           # Add in GenBank data: check name, then add length, description, chromosome etc
           if ($dets[$j][0] eq $getname[1]) {
        print OUT2 "\t$getname[2]\t"; # print length
               for ($zzz = $defin; $zzz < $accessn; $zzz++) {
                   if (!(($getname[$zzz] =~ /allus/) || ($getname[$zzz] =~ /RNA/))) {
                      print OUT2 "$getname[$zzz] "; } } # print description
               if ($chrom == 999) { print OUT2 "\tchromosome=Unknown"; }
               else { $getname[$chrom] =~ s/\///g;
                      $getname[$chrom] =~ s/\"//g; # print chromosome
                      print OUT2 "\t$getname[$chrom]"; }

               $getname[$genename] =~ s/\"//g;
 # print gene short name
               $getname[$genename] =~ s/gene=//g;
               $getname[$genename] =~ s/\///g;
               print OUT2 "\t$getname[$genename]";            }       }   } }

#print "\n\nCheck: number of GenBank entry names in total is ";
#system("grep -c 'M_' temp1");
#print "\nCheck: number of GenBank entry names with SNPs is ";
#system("grep -c 'M_' snphit");
# Export details back to MySQL table (SNPhits)
exit;
```

Appendix B – EST SNP Gene Database

All 1,296 genes identified with $P_N/P_S > 0.5$ or non-con/con ≥ 0.5.

The genes are unranked. [1] Total number of SNPs in ESTs observed at the locus. P_N/P_S or non-con/con is defined as "high" where $P_S = 0$ and $P_N > 0$ or con = 0 and non-con > 0, respectively. Non-con stands for non-conservative replacement substitutions and con for conservative ones.

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
hypothetical protein FLJ3287	XM_414934	1	2	0.500	1	0	high	4
LOC422953 LOC422953	XM_430006	1	0	high	1	0	high	2
serpin peptidase inhibitor clade B ovalbumin	NM_001006377	1	1	1.000	0	1	0.000	3
LOC420040 LOC420040	XM_429727	1	0	high	0	1	0.000	2
tropomodulin 3 LOC415421	XM_413804	2	4	0.500	0	2	0.000	7
Dehydrogenase/reductase SDR	XM_421423	1	0	high	0	1	0.000	2
rai-like protein LOC430363	XM_427923	1	1	1.000	1	0	high	3
Nop132 LOC415955	XM_414299	2	0	high	2	0	high	3
crystallin beta B3 CRYBB3	NM_205191	2	1	2.000	0	2	0.000	4
ENSANGP00000017034 LOC42322	XM_421148	3	1	3.000	2	1	2.000	5
NECAP endocytosis associated 2 NECAP2	NM_001012837	8	2	4.000	5	3	1.667	11
hypothetical protein LOC2188	XM_421523	1	0	high	0	1	0.000	2
hypothetical protein FLJ1008	XM_424221	1	1	1.000	0	1	0.000	3
hypothetical protein FLJ1451	XM_419083	2	1	2.000	0	2	0.000	4
Na+-glucose cotransporter ty	XM_414862	1	0	high	1	0	high	2
60S ribosomal protein L19 m	XM_420071	1	1	1.000	0	1	0.000	3
matrix metalloproteinase 23B	XM_417569	1	1	1.000	0	1	0.000	3
NAD kinase NADK	NM_001030870	1	0	high	0	1	0.000	2
unc-93 homolog A; unc93	XM_419606	1	0	high	0	1	0.000	2
PTK9L protein tyrosine kinase 9-like A6-related	NM_001030589	1	0	high	0	1	0.000	2
cytochrome P450 2H1 CYP2H1	NM_001001616	1	1	1.000	0	1	0.000	3
KIAA1181 protein LOC416205	XM_414530	3	0	high	1	2	0.500	4
ribosomal protein P0-like pr	XM_425751	11	4	2.750	5	6	0.833	16
potassium large conductance calcium-activated chan	NM_204602	1	0	high	0	1	0.000	2
KIAA0852 protein LOC416994	XM_415286	3	0	high	0	3	0.000	4
phosphatidylinositol transfer protein beta PITPN	NM_001039266	2	3	0.667	0	2	0.000	6
Hypothetical UPF0193 protein	XM_416268	2	1	2.000	0	2	0.000	4
amyotrophic lateral sclerosi	XM_421938	1	0	high	0	1	0.000	2
SCY1-like 3 S. cerevisiae SCYL3	NM_001012595	1	1	1.000	0	1	0.000	3
zinc finger FYVE domain containing 27 ZFYVE27	NM_001039304	3	0	high	0	3	0.000	4
DSCR1-like protein LOC41851	XM_416719	1	1	1.000	0	1	0.000	3
XIAP associated factor-1 iso	XM_415922	1	0	high	0	1	0.000	2
acyl-Coenzyme A binding domain containing 5 ACBD5	NM_001006356	2	0	high	1	1	1.000	3
Ubiquitin carboxyl-terminal hydrolase 1	NM_001031290	5	3	1.667	1	4	0.250	9
ring finger protein 7 RNF7	NM_001031307	2	1	2.000	2	0	high	4
beta-adrenergic-receptor kin	XM_415195	1	0	high	1	0	high	2
TBC1 domain family member 2	XM_424947	1	0	high	0	1	0.000	2
82-kD FMRP Interacting Prote	XM_415828	1	0	high	0	1	0.000	2
adiponectin receptor 1 ADIPOR1	NM_001031027	1	2	0.500	0	1	0.000	4
Ku p70/p80 protein LOC424	XM_422072	1	0	high	1	0	high	2
ubiquitin fusion degradation 1 like yeast UFD1L	NM_204301	1	1	1.000	0	1	0.000	3
hypothetical protein FLJ2008	XM_423348	1	1	1.000	0	1	0.000	3
acetyl-Coenzyme A acyltransferase 2 mitochondrial	NM_001006571	2	0	high	0	2	0.000	3
RIKEN cDNA 4933428G09 LOC42	XM_420764	1	1	1.000	0	1	0.000	3
hexokinase 1 HK1	NM_204101	2	4	0.500	1	1	1.000	7
hypothetical protein FLJ1072	XM_417931	1	0	high	0	1	0.000	2
pericentriolar materiall PCM1	NM_204531	2	0	high	0	2	0.000	3
NADH dehydrogenase ubiquinone 1 alpha subcomplex	NM_001006281	1	1	1.000	0	1	0.000	3
hypothetical protein DKFZp43	XM_414936	2	3	0.667	1	1	1.000	6
GLP_39_88222_87572 LOC41794	XM_416190	1	0	high	1	0	high	2

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
Chromosome 10 open reading frame	XM_421536	1	0	high	0	1	0.000	2
KIAA1580 protein LOC428862	XM_426419	2	0	high	1	1	1.000	3
Chondrolectin precursor Tra	XM_416682	1	0	high	0	1	0.000	2
hypothetical protein FLJ2053	XM_423489	3	1	3.000	0	3	0.000	5
chromosome 17 open reading frame	XM_420083	1	0	high	0	1	0.000	2
RIKEN cDNA 2810039F03 LOC41	XM_415020	1	2	0.500	0	1	0.000	4
hypothetical protein MGC1552	XM_420074	2	0	high	0	2	0.000	3
thyroid hormone receptor	XM_413717	2	0	high	1	1	1.000	3
ribosomal protein L29	XM_425143	5	0	high	4	1	4.000	6
ceroid-lipofuscinosis neuronal 8 epilepsy	NM_001031087	1	1	1.000	0	1	0.000	3
Proteasome subunit beta type	XM_417777	2	4	0.500	2	0	high	7
Hypothetical protein KIAA010	XM_413716	2	0	high	0	2	0.000	3
seven-pass transmembrane rec	XM_423746	1	1	1.000	0	1	0.000	3
NADH-ubiquinone oxidoreducta	XM_427322	2	2	1.000	0	2	0.000	5
RIKEN cDNA 1110018M03 LOC42	XM_424544	1	0	high	1	0	high	2
mitochondrial ribosomal prot	XM_420108	3	3	1.000	2	1	2.000	7
putative bHLH transcription	XM_420985	1	1	1.000	0	1	0.000	3
Serine or cysteine protein	XM_421341	1	0	high	1	0	high	2
SH3-domain GRB2-like endophilin B1 SH3GLB1	NM_001006534	2	0	high	1	1	1.000	3
RIKEN cDNA 3110009E18 LOC42	XM_422121	1	1	1.000	0	1	0.000	3
survival of motor neuron protein interacting protein	NM_001039302	1	1	1.000	0	1	0.000	3
protocadherin 10 PCDH10	NM_214672	1	0	high	0	1	0.000	2
transmembrane channel-like 5	XM_414913	1	0	high	0	1	0.000	2
Potential phospholipid-trans	XM_416881	2	0	high	0	2	0.000	3
bumetanide-sensitive Na-K-Cl	XM_413814	1	2	0.500	0	1	0.000	4
Hypothetical protein FLJ1244	XM_415259	2	3	0.667	0	2	0.000	6
prepro vasoactive intestinal	XM_417707	1	2	0.500	0	1	0.000	4
protein-L-isoaspartate D-aspartate O-methyltrans	NM_001031525	1	1	1.000	0	1	0.000	3
v-myb myeloblastosis viral oncogene homolog avian	NM_205318	1	1	1.000	1	0	high	3
N-acetylneuraminic acid synthase sialic acid synt	NM_001007975	1	1	1.000	0	1	0.000	3
defender against cell death 1 DAD1	NM_001007473	1	0	high	0	1	0.000	2
TAF3 protein LOC419107	NM_001030841	2	1	2.000	1	1	1.000	4
ribosomal protein S4 LOC396001	NM_205108	9	10	0.900	4	5	0.800	20
immature colon carcinoma tra	XM_420117	1	0	high	0	1	0.000	2
hypothetical protein DKFZp76	XM_420574	1	0	high	0	1	0.000	2
Mitochondrial ribosomal prot	XM_415911	3	1	3.000	0	3	0.000	5
chromosome X open reading fr	XM_416789	1	1	1.000	1	0	high	3
expressed sequence AW260253	XM_420338	1	0	high	0	1	0.000	2
proteasome prosome macropain 26S subunit non-A	NM_001031256	1	1	1.000	1	0	high	3
Jumonji domain containing pr	XM_422410	1	1	1.000	1	0	high	3
cell division cycle 73 polymerase II complex comp	NM_001031265	1	2	0.500	1	0	high	4
K60 protein LOC422654	XM_420608	1	0	high	0	1	0.000	2
bA207C16.2 novel protein	XM_424813	2	0	high	1	1	1.000	3
B-cell CLL/lymphoma 7A	XM_415148	3	0	high	3	0	high	4
Argininosuccinate synthase Citrulline-	NM_001013395	1	1	1.000	0	1	0.000	3
RIKEN cDNA 0610011N22 LOC420970	NM_001031006	1	0	high	0	1	0.000	2
guanylin precursor LOC41949	XM_417652	2	1	2.000	0	2	0.000	4
liver expressed antimicrobial peptide 2 LOC414338	NM_001001606	1	0	high	1	0	high	2
hypothetical protein LOC1999	XM_422496	2	3	0.667	0	2	0.000	6
hypothetical protein FLJ20457 LOC42098	NM_001006383	2	1	2.000	1	1	1.000	4
Type I iodothyronine deiodinase LOC395	XM_422487	1	0	high	0	1	0.000	2
IL2 alpha receptor LOC395294	NM_204596	1	1	1.000	0	1	0.000	3
KIAA1571 protein LOC424100	XM_421954	1	0	high	0	1	0.000	2
Williams Beuren syndrome chromosome region 22 WBS	NM_001039332	1	2	0.500	0	1	0.000	4
MYST histone acetyltransferase 2	NM_001031341	1	2	0.500	1	0	high	4
RIKEN cDNA 2410021P16 LOC41	XM_415170	2	0	high	0	2	0.000	3
Desmoplakin DP 250/210 kD	XM_418957	4	0	high	1	3	0.333	5
protein tyrosine phosphatase non-receptor type 6	NM_001031484	2	0	high	0	2	0.000	3
acetyl-Coenzyme A acetyltran	XM_417162	1	1	1.000	0	1	0.000	3
chromosome 10 open reading frame	XM_417230	1	0	high	1	0	high	2
integrin beta 1 fibronectin receptor beta polypeptide	NM_001039254	1	2	0.500	1	0	high	4
bA305P22.2.1 novel protein isoform 1	NM_001006292	1	1	1.000	1	0	high	3

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
B6.1 LOC396098	NM_205182	6	3	2.000	4	2	2.000	10
HCV NS3-transactivated protein	XM_424423	1	0	high	1	0	high	2
retinoic acid receptor responder tazarotene	NM_204534	7	2	3.500	2	5	0.400	10
LOC424630 LOC424630	XM_430168	1	0	high	0	1	0.000	2
WD repeat domain 3 WDR3	NM_001031485	1	0	high	1	0	high	2
transient receptor potential cation channel	NM_204692	2	1	2.000	0	2	0.000	4
coronin 7 CORO7	NM_001006176	3	5	0.600	0	3	0.000	9
CD81 antigen target of antiproliferative antibody	NM_001030339	2	0	high	0	2	0.000	3
forkhead box M1 FOXM1	NM_001012955	1	0	high	0	1	0.000	2
RIKEN cDNA A830094I09 gene	XM_420005	1	0	high	1	0	high	2
beta-defensin 10 GAL10	NM_001001609	1	0	high	1	0	high	2
hypothetical protein FLJ1083	XM_421573	1	0	high	1	0	high	2
FKSG26 protein LOC425450	XM_423211	1	0	high	0	1	0.000	2
interleukin 25; lymphocyte antigen 6 co	NM_001006342	1	0	high	0	1	0.000	2
hect domain and RLD 5	XM_420476	1	0	high	0	1	0.000	2
mouse double minute 4	XM_417957	1	0	high	1	0	high	2
aquaporin 9 LOC415402	XM_413787	1	2	0.500	1	0	high	4
Enhancer of zeste homolog 2	XM_418879	1	1	1.000	0	1	0.000	3
superkiller viralicidic activity 2-like 2	NM_001012944	1	1	1.000	0	1	0.000	3
RIKEN cDNA A630033H20 gene	XM_420148	2	1	2.000	0	2	0.000	4
ribosomal protein L7 RPL7	NM_001006345	12	4	3.000	3	9	0.333	17
2'-5'-oligoadenylate synthetase-like OASL	NM_205041	2	2	1.000	1	1	1.000	5
Butyrophilin precursor BT	XM_424442	2	4	0.500	2	0	high	7
Hypothetical protein MGC7567	XM_424043	1	1	1.000	1	0	high	3
phosphorylase kinase gamma 1 muscle PHKG1	NM_001006217	3	6	0.500	0	3	0.000	10
exocyst complex component 7 EXOC7	NM_001012802	1	1	1.000	1	0	high	3
KIAA0837 protein LOC416324	XM_414641	1	1	1.000	0	1	0.000	3
actinin alpha 2 ACTN2	NM_205323	2	2	1.000	1	1	1.000	5
Catalase LOC423601 partial	XM_421487	4	1	4.000	2	2	1.000	6
MCM2 minichromosome maintenance deficient 2 mitot	NM_001006139	1	0	high	1	0	high	2
proteasome prosome macropain subunit beta type	NM_001007905	1	1	1.000	1	0	high	3
hypothetical gene supported by BX931271	XM_417971	3	0	high	2	1	2.000	4
EBNA1 binding protein 2 LOC	XM_422396	2	3	0.667	0	2	0.000	6
transmembrane protein 59 TMEM59	NM_001006541	1	0	high	0	1	0.000	2
acyl-Coenzyme A binding domain containing 3 ACBD3	NM_001031043	1	2	0.500	0	1	0.000	4
MGC68817 protein LOC416456	XM_414764	1	1	1.000	1	0	high	3
zinc finger protein 291 LOC	XM_413735	1	0	high	0	1	0.000	2
methylenetetrahydrofolate dehydrogenase NADP+	NM_001031360	1	0	high	0	1	0.000	2
aldehyde oxidase 2 LOC424072	NM_001039601	1	1	1.000	0	1	0.000	3
transglutaminase y LOC41921	XM_417393	1	1	1.000	0	1	0.000	3
Polo-like kinase 4 LOC42249	XM_420462	1	0	high	0	1	0.000	2
D-dopachrome tautomerase DDT	NM_001030667	1	1	1.000	0	1	0.000	3
membrane-associated protein	XM_422456	1	0	high	0	1	0.000	2
hypothetical protein LOC424	XM_422119	2	2	1.000	0	2	0.000	5
NADH dehydrogenase ubiquinone	XM_421103	2	3	0.667	0	2	0.000	6
dihydropyrimidinase LOC4202	XM_418377	5	2	2.500	1	4	0.250	8
chemokine C-X-C motif receptor 4 CXCR4	NM_204617	2	0	high	1	1	1.000	3
phosphatidylinositol glycan	XM_423199	1	0	high	1	0	high	2
cell cycle progression 1 CCPG1	NM_001031455	2	1	2.000	1	1	1.000	4
Zinc finger protein ZFPM2 Z	XM_418379	1	0	high	1	0	high	2
syndecan 4 amphiglycan ryudocan SDC4	NM_001007869	1	0	high	0	1	0.000	2
NAD-dependent deacetylase SIRT2 SIRT2	NM_001017414	5	3	1.667	2	3	0.667	9
protease serine 2 trypsin 2 PRSS2	NM_205384	18	11	1.636	10	8	1.250	30
4933434G05Rik protein LOC41	XM_415636	2	2	1.000	2	0	high	5
Adenomatous polyposis coli	XM_413975	1	0	high	1	0	high	2
adaptor-related protein complex 1 sigma 2 subunit	NM_001006261	2	0	high	0	2	0.000	3
modulator recognition factor	XM_428598	1	0	high	0	1	0.000	2
ADP-ribosylhydrolase like 2 ADPRHL2	NM_001006312	2	0	high	0	2	0.000	3
C-type lectin domain family 3 member B CLEC3B	NM_204666	4	0	high	1	3	0.333	5
pinopsin LOC396377	NM_205409	1	1	1.000	1	0	high	3
myosin heavy polypeptide 6 cardiac muscle alpha	NM_204766	1	0	high	0	1	0.000	2
Biliverdin reductase A precu	XM_418872	1	2	0.500	1	0	high	4

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
Vaccinia related kinase 3 L	XM_415042	1	0	high	0	1	0.000	2
Probable cation-transporting	XM_423767	1	2	0.500	0	1	0.000	4
hypothetical protein BC011840 LOC42608	NM_001031355	4	1	4.000	2	2	1.000	6
integrin beta 5 ITGB5	NM_204483	1	2	0.500	1	0	high	4
Ribonucleoside-diphosphate	XM_419948	1	1	1.000	0	1	0.000	3
PHD finger protein 3 LOC428	XM_426199	2	1	2.000	1	1	1.000	4
proteasome prosome macropain 26S subunit ATPas	NM_001031190	1	2	0.500	1	0	high	4
mKIAA1453 protein LOC426457	XM_424101	2	0	high	1	1	1.000	3
r-goliath LOC416281	XM_414601	2	3	0.667	0	2	0.000	6
Stromal cell-derived factor	XM_425406	4	4	1.000	0	4	0.000	9
ras-related C3 botulinum toxin substrate 2 rho fa	NM_001012536	1	0	high	0	1	0.000	2
NDP52 LOC419993	XM_418115	2	2	1.000	0	2	0.000	5
solute carrier family 25 mitochondrial carrier b	NM_001012883	1	2	0.500	0	1	0.000	4
Vesicle amine transport prot	XM_418130	1	0	high	1	0	high	2
myosin light polypeptide kinase MYLK	NM_205459	1	1	1.000	0	1	0.000	3
fumarate hydratase FH	NM_001006382	1	1	1.000	1	0	high	3
Proteasome subunit alpha typ	XM_413742	1	2	0.500	0	1	0.000	4
hypothetical protein FLJ2386	XM_421860	1	1	1.000	0	1	0.000	3
annexin VIII; VAC beta LOC4	XM_421646	1	1	1.000	0	1	0.000	3
KIAA0542 protein LOC427706	XM_425281	1	1	1.000	1	0	high	3
ebip7226-like LOC395166	XM_420617	1	2	0.500	0	1	0.000	4
erythrocyte membrane protein	XM_424362	1	1	1.000	0	1	0.000	3
hypothetical protein LOC423	XM_421853	1	0	high	1	0	high	2
family with sequence similarity 53 member A FAM5	NM_204388	1	2	0.500	1	0	high	4
v-jun sarcoma virus 17 oncogene homolog avian J	NM_001031289	2	1	2.000	0	2	0.000	4
hypothetical protein FLJ1176	XM_418657	1	1	1.000	0	1	0.000	3
apical early endosomal glyco	XM_418614	1	0	high	1	0	high	2
Caldecrin precursor Chymotr	XM_428223	5	4	1.250	2	3	0.667	10
uncoupling protein 3 mitochondrial proton carrie	NM_204107	1	0	high	0	1	0.000	2
malignant T cell amplified	XM_420334	1	0	high	0	1	0.000	2
zinc finger A20 domain containing 2 ZA20D2	NM_001031424	3	2	1.500	0	3	0.000	6
RIKEN cDNA 1810030N24 LOC42	XM_419851	2	1	2.000	1	1	1.000	4
Secretogranin I precursor S	XM_419377	3	3	1.000	0	3	0.000	7
P311 POU 3.1	NM_205391	1	0	high	0	1	0.000	2
t-complex testis expressed 1	XM_419699	1	1	1.000	1	0	high	3
hypothetical gene supported by NM_20506	XM_429275	3	3	1.000	0	3	0.000	7
fibronectin type III domain containing 3A FNDC3A	NM_001012826	1	1	1.000	0	1	0.000	3
copine I LOC419134	XM_417319	2	3	0.667	1	1	1.000	6
ribosomal protein S24 LOC42	XM_421602	2	0	high	0	2	0.000	3
probable ABC-type transport	XM_425303	1	1	1.000	1	0	high	3
LOC421970 LOC421970	XM_429903	4	4	1.000	1	3	0.333	9
guanine deaminase LOC427253	XM_424835	1	0	high	0	1	0.000	2
prostaglandin D2 synthase 21kDa brain PTGDS	NM_204259	1	2	0.500	0	1	0.000	4
anti-Mullerian hormone AMH	NM_205030	1	0	high	1	0	high	2
DnaJ Hsp40 homolog subfamily A member 1 DNAJA	NM_001012945	1	0	high	0	1	0.000	2
OTU domain containing 6B OTUD6B	NM_001006347	2	4	0.500	1	1	1.000	7
histone macroH2A1.2 LOC395858	NM_205007	2	1	2.000	0	2	0.000	4
enolase LOC396016	NM_205119	2	2	1.000	1	1	1.000	5
DPH1 homolog S. cerevisiae DPH1	NM_001030716	1	0	high	0	1	0.000	2
von Willebrand factor LOC41	XM_417223	1	0	high	0	1	0.000	2
phosphatidylserine synthase 1 PTDSS1	NM_001031505	1	0	high	0	1	0.000	2
apolipoprotein H precursor	XM_415683	2	2	1.000	0	2	0.000	5
mitochondrial ribosomal protein	XM_419495	2	2	1.000	1	1	1.000	5
staufen binding protein homolog 2 Drosophila	NM_001030941	2	4	0.500	0	2	0.000	7
LOC417337 LOC417337	XM_429497	1	1	1.000	0	1	0.000	3
mitochondrial ribosomal prot	XM_419444	1	2	0.500	0	1	0.000	4
peroxiredoxin 1 LOC424598	XM_422437	7	6	1.167	2	5	0.400	14
normal mucosa of esophagus	XM_413822	1	0	high	0	1	0.000	2
KIAA1474 protein LOC415762	XM_414125	1	1	1.000	0	1	0.000	3
golgi SNAP receptor complex member 1 GOSR1	NM_001006222	1	1	1.000	0	1	0.000	3
LOC419161 LOC419161	XM_429631	1	0	high	1	0	high	2
activating transcription factor 4 tax-responsive	NM_204880	1	0	high	0	1	0.000	2

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
5'-nucleotidase cytosolic II NT5C2	NM_001031234	1	0	high	0	1	0.000	2
Rhesus blood group-associated glycoprotein RHAG	NM_204464	2	1	2.000	0	2	0.000	4
hypothetical protein FLJ21908 LOC41781	NM_001006231	1	0	high	0	1	0.000	2
LYRIC LOC420239	XM_418351	1	0	high	0	1	0.000	2
Hypothetical protein KIAA0258 LOC43165	NM_001031611	1	0	high	0	1	0.000	2
immune costimulatory protein	XM_416760	1	0	high	0	1	0.000	2
COP9 constitutive photomorphogenic homolog subunit	NM_001031596	2	3	0.667	1	1	1.000	6
fibronectin type 3 and SPRY	XM_428164	1	0	high	0	1	0.000	2
Myosin regulatory light chain	XM_415166	1	2	0.500	1	0	high	4
Thioredoxin-like protein 2	XM_421826	2	4	0.500	0	2	0.000	7
hypothetical protein LOC421	XM_419286	1	0	high	1	0	high	2
oxidative stress responsive 1 LOC419203	NM_001029981	3	6	0.500	0	3	0.000	10
proteasome prosome macropain 26S subunit non-A	NM_001006189	1	0	high	1	0	high	2
pyruvate dehydrogenase kinase isozyme 1 PDK1	NM_001031352	1	1	1.000	1	0	high	3
phosphoinositide-3-kinase catalytic delta polype	NM_001012696	2	3	0.667	0	2	0.000	6
casein kinase 1 alpha 1 CSNK1A1	NM_205053	1	0	high	1	0	high	2
phosphatidylinositol-4-phosphate 5-kinase type II	NM_001030971	1	2	0.500	0	1	0.000	4
zinc finger DHHC-type containing 17 ZDHHC17	NM_001030745	1	0	high	0	1	0.000	2
mitochondrial ribosomal protein L48 MRPL48 nucl	NM_001030053	1	0	high	0	1	0.000	2
hypothetical protein FLJ11301 LOC42489	NM_001006549	2	4	0.500	0	2	0.000	7
sorting nexin 10 SNX10	NM_001030986	2	3	0.667	2	0	high	6
cytochrome c oxidase subunit	XM_419450	1	2	0.500	0	1	0.000	4
high-mobility group 20A HMG20A	NM_001030394	1	0	high	1	0	high	2
Retinol-binding protein II	XM_422636	2	0	high	0	2	0.000	3
heat shock transcription	XM_416754	1	1	1.000	0	1	0.000	3
protein phosphatase 2 regulatory subunit B delta	NM_001006507	1	0	high	0	1	0.000	2
helicase MOV-10-like LOC419872	NM_001012843	1	1	1.000	1	0	high	3
pantothenate kinase 4	XM_417556	1	0	high	1	0	high	2
cytosolic ovarian carcinoma antigen 1 COVA1	NM_001006427	3	0	high	1	2	0.500	4
RIKEN cDNA 2400003L07 LOC42	XM_422733	3	2	1.500	2	1	2.000	6
RER1 homolog LOC419397	NM_001006300	1	0	high	0	1	0.000	2
hypothetical gene supported by CR386985	XM_418889	1	1	1.000	0	1	0.000	3
Exosome complex exonuclease	XM_417092	1	2	0.500	0	1	0.000	4
heme oxygenase decycling 1 HMOX1	NM_205344	2	3	0.667	0	2	0.000	6
fatty acid binding protein 4 adipocyte FABP4	NM_204290	2	2	1.000	0	2	0.000	5
fatty acid amide hydrolase	XM_422450	1	2	0.500	0	1	0.000	4
LOC426122 LOC426122	XM_430363	4	2	2.000	3	1	3.000	7
MYC induced nuclear antigen	XM_416647	2	4	0.500	0	2	0.000	7
matrix metalloproteinase 28	XM_415771	1	0	high	0	1	0.000	2
UbiA prenyltransferase domain containing 1 UBIAD1	NM_001030879	2	2	1.000	0	2	0.000	5
Nucleoporin-like protein RIP	XM_422611	1	0	high	0	1	0.000	2
syntaxin 8 STX8	NM_001030698	1	0	high	1	0	high	2
GOB-4 LOC420596	XM_418698	2	2	1.000	1	1	1.000	5
Zinc finger protein 142 HA4	XM_423456	1	0	high	0	1	0.000	2
cyclin K CCNK	NM_001031209	1	0	high	0	1	0.000	2
exocyst complex component 3 EXOC3	NM_001006384	2	2	1.000	2	0	high	5
3-hydroxyacyl-CoA dehydrogen	XM_423479	2	0	high	2	0	high	3
phosphoglycerate dehydrogenase like 1	NM_001006268	1	0	high	0	1	0.000	2
hypothetical protein FLJ22626 LOC42236	NM_001006437	3	0	high	2	1	2.000	4
K60 protein K60	NM_205018	3	1	3.000	1	2	0.500	5
thymine-DNA glycosylase TDG	NM_204750	1	1	1.000	0	1	0.000	3
family with sequence similar	XM_415873	1	2	0.500	1	0	high	4
hypothetical gene supported by CR353961	XM_429985	6	1	6.000	4	2	2.000	8
matrix Gla protein MGP	NM_205044	4	0	high	0	4	0.000	5
mitochondrial ribosomal prot	XM_413871	1	0	high	0	1	0.000	2
fructose-16-bisphosphatase FBP1	XM_425040	2	4	0.500	1	1	1.000	7
hypothetical gene supported by BX933825	XM_429587	3	2	1.500	2	1	2.000	6
protease serine 11	XM_424004	1	0	high	1	0	high	2
DEAD Asp-Glu-Ala-Asp box polypeptide 27 DDX27	NM_001006293	2	3	0.667	0	2	0.000	6
RIKEN cDNA A930039G15 gene	XM_419399	1	0	high	1	0	high	2
chromosome 11 open reading frame	XM_420984	1	1	1.000	0	1	0.000	3

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
vitellogenin LOC424533	NM_001031276	1	1	1.000	0	1	0.000	3
USP6NL protein LOC419105	XM_417293	1	1	1.000	1	0	high	3
basic transcription factor 3-like 4 BTF3L4	NM_001031285	1	1	1.000	1	0	high	3
Nit protein 2 LOC418386	XM_416604	1	2	0.500	0	1	0.000	4
nicotinate phosphoribosyltransferase domain contai	NM_206981	1	1	1.000	0	1	0.000	3
platelet-activating factor acetylhydrolase isoform	NM_001030911	1	0	high	1	0	high	2
triosephosphate isomerase 1 TPI1	NM_205451	1	2	0.500	0	1	0.000	4
tumor protein p53 inducible nuclear protein 1 TP5	NM_001030946	1	0	high	0	1	0.000	2
NADH-ubiquinone oxidoreducta	XM_414844	1	1	1.000	0	1	0.000	3
RIKEN cDNA 4930521E07 LOC42	XM_421833	3	1	3.000	0	3	0.000	5
glycine cleavage system protein H aminomethyl	NM_001004372	2	3	0.667	1	1	1.000	6
RIKEN cDNA 1110005A03 LOC41	XM_415616	1	1	1.000	0	1	0.000	3
Protein FAM3C precursor Protein	XM_416002	1	2	0.500	0	1	0.000	4
inhibitor of DNA binding 4 dominant negative heli	NM_204282	1	0	high	1	0	high	2
24-dienoyl-CoA reductase	XM_418328	7	4	1.750	2	5	0.400	12
gem nuclear organelle associated protein 4 GEMI	NM_001012610	2	0	high	0	2	0.000	3
hydrolase 2 mitochondrial	XM_415879	3	2	1.500	1	2	0.500	6
brain protein 44-like LOC428592	NM_001031524	1	0	high	0	1	0.000	2
PWP1 homolog S. cerevisiae PWP1	NM_001030761	1	1	1.000	0	1	0.000	3
interleukin 16 lymphocyte chemoattractant factor	NM_204352	5	6	0.833	2	3	0.667	12
small glutamine-rich tetratricopeptide repeat TPR	NM_001031379	2	0	high	1	1	1.000	3
Claudin-12 LOC420545	XM_418646	1	1	1.000	0	1	0.000	3
lipoprotein APOVLDLII	NM_205483	1	2	0.500	0	1	0.000	4
Shmt1-prov protein LOC41652	XM_414824	3	2	1.500	1	2	0.500	6
Saccharomyces cerevisiae Nip	XM_414222	1	2	0.500	0	1	0.000	4
cdk inhibitor CIP1 p21 CIP1	NM_204396	2	0	high	2	0	high	3
Ig mu heavy chain disease	XM_428803	2	0	high	1	1	1.000	3
PIT 54 protein PIT 54	NM_207180	2	1	2.000	0	2	0.000	4
Transmembrane 4 superfamily	XM_416779	1	1	1.000	1	0	high	3
macrothioredoxin LOC424115	XM_421968	2	3	0.667	0	2	0.000	6
tyrosylprotein sulfotransferase 2 TPST2	NM_001012794	1	1	1.000	0	1	0.000	3
Sec61 alpha isoform 2 LOC42	XM_424024	1	0	high	0	1	0.000	2
transthyretin prealbumin amyloidosis type I TT	NM_205335	4	3	1.333	1	3	0.333	8
chromogranin A precursor	XM_421330	1	1	1.000	0	1	0.000	3
gap junction protein alpha 1 43kDa connexin 43	NM_204586	2	1	2.000	0	2	0.000	4
troponin I type 2 skeletal fast TNNI2	NM_205417	2	3	0.667	1	1	1.000	6
NIMA never in mitosis gene a-related kinase 2 N	NM_001031050	1	1	1.000	0	1	0.000	3
amylase alpha 1A; salivary AMY1A	NM_001001473	23	5	4.600	9	1	4 0.643	29
procarboxypeptidase B LOC42	XM_422699	5	4	1.250	0	5	0.000	10
chymotrypsin EC 3.4.21.1 2	XM_428788	4	5	0.800	2	2	1.000	10
Carboxypeptidase M precursor	XM_416085	1	0	high	1	0	high	2
enhancer of rudimentary homolog Drosophila ERH	NM_001006475	1	2	0.500	1	0	high	4
heterogeneous nuclear ribonucleoprotein R HNRPR	NM_001006309	1	0	high	0	1	0.000	2
class II histocompatibility	XM_415339	1	1	1.000	0	1	0.000	3
Hypothetical protein MGC7631	XM_415557	2	1	2.000	1	1	1.000	4
ribosomal protein S23 LOC42	XM_424903	4	0	high	1	3	0.333	5
zinc finger FYVE domain con	XM_421391	1	0	high	0	1	0.000	2
kinesin family member 21A	XM_423537	3	0	high	0	3	0.000	4
karyopherin beta 1; nuclear	XM_424140	3	0	high	0	3	0.000	4
ribosomal protein L26 LOC396400	XM_414531	9	0	high	3	6	0.500	10
mitochondrial ATP synthase	XM_416717	3	2	1.500	2	1	2.000	6
DMRT1 isoform e LOC395181	XM_418817	12	4	3.000	5	7	0.714	17
glyceraldehyde-3-phosphate dehydrogenase GAPDH	NM_204305	14	6	2.333	8	6	1.333	21
CD74 antigen invariant polypeptide of major histo	NM_001001613	3	5	0.600	0	3	0.000	9
CD3E antigen epsilon polypeptide TiT3 complex	NM_206904	8	1	8.000	3	5	0.600	10
brain abundant membrane attached signal protein 1	NM_204116	3	0	high	1	2	0.500	4
non-POU domain containing octamer-binding NONO	NM_001031532	1	0	high	0	1	0.000	2
polyubiquitin LOC417602	XM_415847	15	11	1.364	7	8	0.875	27
sin3-associated polypeptide 18kDa SAP18	NM_204312	2	3	0.667	0	2	0.000	6
40S ribosomal protein S16 L	XM_416113	1	2	0.500	1	0	high	4
Y box binding protein 1 YBX1	NM_204414	9	3	3.000	4	5	0.800	13
RIKEN cDNA 2700055K07 LOC42	XM_424853	1	0	high	1	0	high	2

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
tyrosine 3/tryptophan 5 -monooxygenase	NM_001031343	2	2	1.000	0	2	0.000	5
hypothetical protein FLJ2255	XM_421922	1	1	1.000	0	1	0.000	3
hypothetical gene supported by CR354325	XM_430074	1	0	high	1	0	high	2
guanine nucleotide binding protein G protein be	NM_001004378	5	5	1.000	3	2	1.500	11
a-actin LOC421534	NM_001031063	2	4	0.500	1	1	1.000	7
ribosomal protein L39 RPL39	NM_204272	1	0	high	1	0	high	2
pyruvate kinase muscle	NM_205469	5	5	1.000	1	4	0.250	11
ribosomal protein L7a RPL7A	NM_001004379	17	6	2.833	5	1	2 0.417	24
Feather keratin I Keratin	XM_427209	2	0	high	2	0	high	3
peroxiredoxin 6 PRDX6	NM_001039329	1	0	high	1	0	high	2
myotrophin MTPN	NM_204886	1	0	high	1	0	high	2
protein phosphatase 2 formerly 2A regulatory su	NM_001030886	2	0	high	1	1	1.000	3
ADP-ribosylation factor 1 ARF1	NM_001006352	1	0	high	0	1	0.000	2
eukaryotic translation elongation factor 1 alpha 1	NM_204157	18	5	3.600	8	1	0 0.800	24
proline-rich nuclear receptor coactivator 1 PNRC1	NM_001012291	1	2	0.500	0	1	0.000	4
hematological and neurological expressed 1 HN1	NM_001006425	3	0	high	2	1	2.000	4
protein disulfide isomerase family A member 4 PD	NM_001006370	8	3	2.667	4	4	1.000	12
leucine aminopeptidase 3 LAP3	NM_001031336	2	3	0.667	0	2	0.000	6
beta-H globin LOC428114	NM_001031489	2	3	0.667	0	2	0.000	6
Loc245963-prov protein LOC4	XM_415421	2	2	1.000	1	1	1.000	5
Feather keratin I Keratin	XM_424523	1	2	0.500	0	1	0.000	4
hypothetical gene supported by BX950836	XM_428876	1	0	high	1	0	high	2
hemoglobin delta HBD	NM_205489	6	4	1.500	2	4	0.500	11
v-myc myelocytomatosis viral related oncogene neu	NM_001031091	1	1	1.000	0	1	0.000	3
ubiquitin-52 amino acid fusion protein UBA52	NM_205075	3	5	0.600	1	2	0.500	9
mitochondrial ribosomal protein L23 MR	XM_421027	2	2	1.000	0	2	0.000	5
COP9 constitutive photomorphogenic homolog subunit	NM_001006163	1	0	high	1	0	high	2
40S ribosomal protein S10 L	XM_418029	15	0	high	8	7	1.143	16
scavenger receptor cysteine rich domain containing	NM_001031477	3	5	0.600	2	1	2.000	9
replication protein A232kDa RPA2	NM_001030892	1	2	0.500	1	0	high	4
RAD21 homolog S. pombe RAD21	NM_001030950	2	3	0.667	0	2	0.000	6
heterogeneous nuclear ribonucleoprotein D AU-rich	NM_001031143	1	1	1.000	1	0	high	3
ST3 beta-galactoside alpha-23-sialyltransferase 6	NM_204479	1	0	high	1	0	high	2
actin related protein 2/3	XM_426598	1	0	high	0	1	0.000	2
RIKEN cDNA 2610528K11 LOC417092	NM_001030677	1	1	1.000	0	1	0.000	3
myeloid cellleukemia protein MCL-1 LO	XM_422853	1	0	high	1	0	high	2
early growth response 1(EGR1	NM_204136	1	1	1.000	0	1	0.000	3
Splicing factor arginine/serine-rich 5	NM_001031197	1	1	1.000	0	1	0.000	3
guanine nucleotide binding protein G protein be	NM_001012835	2	0	high	1	1	1.000	3
3-hydroxyisobutyryl-Coenzyme A hydrolase HIBCH	NM_001031243	1	0	high	0	1	0.000	2
ribosomal protein L22 RPL22	NM_204141	3	4	0.750	1	2	0.500	8
ribosomal protein L5 RPL5	NM_204581	3	2	1.500	2	1	2.000	6
heat shock cognate 70 HSC70	NM_205003	17	5	3.400	6	1	1 0.545	23
TRAF and TNF receptor-associ	XM_419093	1	0	high	0	1	0.000	2
ribosomal protein L9; 60S	XM_420741	9	3	3.000	3	6	0.500	13
calmodulin-like protein neoCaM LOC39	XM_421316	1	0	high	0	1	0.000	2
albumin ALB	NM_205261	5	0	high	0	5	0.000	6
trafficking protein particle complex 4 TRAPPC4	NM_001006320	1	0	high	1	0	high	2
aldose 1-epimerase LOC42626	XM_423931	1	2	0.500	0	1	0.000	4
tubulin beta 2B TUBB2B	NM_001004400	2	4	0.500	1	1	1.000	7
apolipoprotein A-I APOA1	NM_205525	19	11	1.727	7	1	2 0.583	31
ferritin heavy polypeptide 1 FTH1	NM_205086	22	5	4.400	13	9	1.444	28
Protein KIAA0494 LOC424617	XM_422454	1	1	1.000	1	0	high	3
chromosome 6open reading frame 51 LOC	NM_001031076	3	4	0.750	1	2	0.500	8
Wpkci WPKCI-8	NM_204688	3	0	high	0	3	0.000	4
cholecystokinin CCK	NM_001001741	2	2	1.000	1	1	1.000	5
heterogeneous nuclear ribonucleoprotein H3 2H9	NM_001012592	2	1	2.000	0	2	0.000	4
eukaryotic translation initi	XM_421787	1	2	0.500	0	1	0.000	4
chromosome 6open reading frame 111	NM_001031080	10	0	high	4	6	0.667	11
ribosomal protein S27a RPS27A	NM_204953	6	2	3.000	2	4	0.500	9
dicarbonyl/L-xylulose reductase DCXR	NM_204225	1	1	1.000	0	1	0.000	3
protein disulfide isomerase family A member 3 PD	NM_204110	2	1	2.000	0	2	0.000	4

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
integrin beta 1 bindingprotein 3 ITGB1BP3	NM_001030550	3	4	0.750	2	1	2.000	8
destrin actin depolymerizing factor DSTN	NM_205528	10	3	3.333	3	7	0.429	14
equarin-L LOC395074	NM_204431	1	2	0.500	0	1	0.000	4
serum/glucocorticoid regulated kinase SGK	NM_204476	1	0	high	0	1	0.000	2
nexin-1 LOC424805	XM_422621	1	2	0.500	0	1	0.000	4
glucose phosphate isomerase GPI	NM_001006128	1	2	0.500	0	1	0.000	4
calcitonin gene-related peptide-recepto	XM_415793	2	3	0.667	0	2	0.000	6
hypothetical gene MGC19595	XM_418230	1	0	high	1	0	high	2
SAM domain and HD domain 1 SAMHD1	NM_001030845	1	0	high	0	1	0.000	2
chromatin modifying protein 2B CHMP2B	NM_001030792	1	0	high	0	1	0.000	2
NADH dehydrogenase ubiquinone 1 beta subcomplex	NM_001006502	1	2	0.500	0	1	0.000	4
SMC5 structural maintenance of chromosomes 5-like	NM_001039335	1	1	1.000	1	0	high	3
lymphocyte antigen 6 complex locus E LY6E	NM_204775	1	1	1.000	0	1	0.000	3
collagen type VI alpha 1 COL6A1	NM_205107	1	0	high	0	1	0.000	2
chaperonin containing TCP1 subunit 6A zeta 1 C	NM_001006216	1	2	0.500	1	0	high	4
tripartite motif-containing 59 TRIM59	NM_001031320	1	2	0.500	0	1	0.000	4
pseudouridine synthase A Ps	XM_415090	1	0	high	1	0	high	2
heterogeneous nuclear ribonucleoprotein A3 HNRPA3	NM_001031253	3	0	high	1	2	0.500	4
Valacyclovir hydrolase precu	XM_418972	1	1	1.000	0	1	0.000	3
Rho GTPase activating protein 26 ARHGAP26	NM_205194	1	0	high	1	0	high	2
hypotheticalprotein 4933408F15 LOC422	NM_001031145	1	1	1.000	0	1	0.000	3
40S ribosomal protein S8 LO	XM_422423	6	3	2.000	4	2	2.000	10
nucleobindin 2 NUCB2	NM_001006468	3	4	0.750	1	2	0.500	8
coiled-colLhelix-coiled-coil	XM_424675	1	0	high	1	0	high	2
HSPC270 LOC416540	XM_414841	2	0	high	0	2	0.000	3
C79952 protein LOC415534	XM_413903	1	1	1.000	0	1	0.000	3
Ectonucleotide pyrophosphata	XM_418466	1	1	1.000	1	0	high	3
interleukin-1 receptor-associated kinase 2 IRAK2	NM_001030605	7	0	high	1	6	0.167	8
pre-fibrinogen alpha subunit LOC396307	NM_205356	2	0	high	0	2	0.000	3
BWK-1 LOC419235	NM_001007871	1	0	high	1	0	high	2
28S ribosomal protein S18c	XM_420566	1	0	high	0	1	0.000	2
collagen type XVIII alpha 1 COL18A1	NM_204164	1	1	1.000	1	0	high	3
eukaryotic translation initiation factor 4A isofo	NM_204549	5	2	2.500	2	3	0.667	8
Elastase 1 precursor LOC425	XM_422984	3	4	0.750	0	3	0.000	8
hypothetical gene supported by CR390807 LOC418765	NM_001006267	3	2	1.500	2	1	2.000	6
Serine/threonine kinase 3 STE20 homolo	NM_001031337	1	0	high	0	1	0.000	2
KIAA0776 protein LOC421804	XM_419830	2	0	high	1	1	1.000	3
ovalbumin LOC396058	NM_205152	2	2	1.000	1	1	1.000	5
hypothetical protein FLJ2231	XM_418840	1	2	0.500	0	1	0.000	4
CG32112-PB LOC422499	NM_001031134	1	0	high	0	1	0.000	2
SGT1 suppressor of G2 allele of SKP1 S. cerevisiae	NM_001030823	2	2	1.000	0	2	0.000	5
peptidylprolyl isomerase cyclophilin-like 2 PPI	NM_001030653	1	0	high	0	1	0.000	2
glutathione-S-transferase	XM_421747	3	4	0.750	0	3	0.000	8
LOC423495 LOC423495	XM_430058	2	1	2.000	1	1	1.000	4
testis-specific leucine zipp	XM_417043	1	0	high	0	1	0.000	2
alcohol dehydrogenase 5(class III chi polypepti	NM_001031152	1	2	0.500	0	1	0.000	4
aminopeptidase-like 1 LOC41	XM_417486	1	0	high	1	0	high	2
muscleblind-like Drosophila MBNL1	NM_001031322	3	2	1.500	2	1	2.000	6
Ndrg3 protein LOC419130	XM_417315	1	0	high	1	0	high	2
Hypotheticalprotein MGC75902 LOC42486	NM_001006544	1	1	1.000	1	0	high	3
ornithine decarboxylaseantizyme 1 OAZ1	NM_204916	6	0	high	4	2	2.000	7
proteasome prosome macropain 26S subunit non-A	NM_001030706	1	1	1.000	1	0	high	3
quiescent cell proline dipep	XM_415570	1	1	1.000	0	1	0.000	3
hypothetical protein FLJ2053	XM_420399	3	1	3.000	1	2	0.500	5
kelch-like 6 Drosophila KLHL6	NM_001031303	1	0	high	0	1	0.000	2
immunoglobulin mu binding protein 2 IGHMBP2	NM_001031175	2	3	0.667	0	2	0.000	6
Phospholipid scramblase 1	XM_422696	1	1	1.000	0	1	0.000	3
suppressor of fused homolog Drosophila SUFU	NM_204264	1	0	high	0	1	0.000	2
leukocyte ribonuclease A-1 LOC396194	NM_205259	2	2	1.000	1	1	1.000	5
leukocyte ribonuclease A-2 RSFR	NM_001007942	3	0	high	1	2	0.500	4
LOC420108 LOC420108	XM_429734	2	2	1.000	1	1	1.000	5

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
RIKEN cDNA A630050E13 gene	XM_419336	1	2	0.500	0	1	0.000	4
gelsolin amyloidosis Finnish type GSN	NM_204934	4	5	0.800	1	3	0.333	10
DNA polymerase-transactivated protein 6	NM_001008474	1	2	0.500	1	0	high	4
Cg14997-prov protein LOC415	XM_413825	1	1	1.000	0	1	0.000	3
phosphatidylinositol 4-kinase type 2 beta PI4K2B	NM_001031157	1	0	high	0	1	0.000	2
hypotheticalgene supported by BX933436	XM_429924	4	2	2.000	3	1	3.000	7
ubiquitin-like containing	XM_418269	1	0	high	1	0	high	2
Hepatocellular carcinoma-associated ant	NM_001006207	1	1	1.000	1	0	high	3
helicase DNA B; helicase B	XM_416077	1	0	high	1	0	high	2
dpy-30-like protein LOC4214	XM_419530	2	2	1.000	0	2	0.000	5
N6-DNA methyltransferase A	XM_416689	3	4	0.750	1	2	0.500	8
hydroxysteroid 17-beta dehydrogenase 4 HSD17B4	NM_204943	5	3	1.667	1	4	0.250	9
2610030H06Rik protein LOC422385	NM_001031127	1	0	high	0	1	0.000	2
phosphoribosyl pyrophosphate synthetase-associated	NM_001006165	1	0	high	1	0	high	2
RAD52 motif 1 RDM1	NM_204546	2	0	high	1	1	1.000	3
phosphatidylethanolamine N-methyltransferase PEMT	NM_001006164	3	2	1.500	1	2	0.500	6
stathmin 1/oncoprotein 18 STMN1	NM_001001858	5	3	1.667	3	2	1.500	9
SLC35E3 protein LOC417842	XM_416083	1	0	high	1	0	high	2
thioesterase B LOC415786	XM_414147	1	0	high	1	0	high	2
hepatoma-derived growth fact	XM_413841	1	1	1.000	0	1	0.000	3
geminin DNA replication inhibitor GMNN	NM_001031010	1	2	0.500	1	0	high	4
testis specific 14; testis	XM_414980	3	3	1.000	0	3	0.000	7
proteasome prosome macropain subunit beta type	NM_204397	4	3	1.333	1	3	0.333	8
FLJ11712 protein LOC418874	NM_001030826	1	1	1.000	0	1	0.000	3
muted homolog mouse MUTED	NM_001006373	1	0	high	0	1	0.000	2
RIKEN cDNA 2900055D14 LOC42	XM_423984	1	0	high	0	1	0.000	2
aminoacylasefamily member 45.2 kD 4	NM_001030915	1	0	high	0	1	0.000	2
G-2 and S-phase expressed 1 GTSE1	NM_001031332	1	1	1.000	0	1	0.000	3
Protein KIAA0179 LOC425237	NM_001031334	2	1	2.000	0	2	0.000	4
protein phosphatase 1E	XM_415871	2	0	high	1	1	1.000	3
RIKEN cDNA D030060M11 LOC428248	NM_001039319	1	1	1.000	1	0	high	3
eukaryotic translation	XM_415738	1	1	1.000	1	0	high	3
diazepam binding inhibitor GABA receptor modulato	NM_204576	1	0	high	0	1	0.000	2
hypotheticalgene supported by CR385540	XM_429479	1	2	0.500	0	1	0.000	4
Ras suppressor protein 1	XM_418627	2	1	2.000	1	1	1.000	4
Protein C21orf59 LOC418497	NM_001006258	2	2	1.000	0	2	0.000	5
fibroblast growth factor receptor 2 bacteria-expr	NM_205319	2	2	1.000	1	1	1.000	5
DPCD protein LOC423852	XM_421721	4	8	0.500	2	2	1.000	13
nucleoporin like 2 NUPL2	NM_001030984	1	1	1.000	0	1	0.000	3
Hypothetical protein MGC6629	XM_422896	1	2	0.500	0	1	0.000	4
cdc2/CDC28-like protein kina	XM_414614	1	1	1.000	0	1	0.000	3
cell cycle progression 2 protein	XM_418515	3	4	0.750	2	1	2.000	8
outer dense fiber of sperm tails 2 ODF2	NM_001012799	1	0	high	1	0	high	2
tyrosine 3-monooxygenase/tryptophan 5-monooxygenas	NM_001006289	2	3	0.667	1	1	1.000	6
Bystin LOC419927	XM_418047	1	0	high	1	0	high	2
bromodomain containing 7 BRD7	NM_001005839	1	0	high	0	1	0.000	2
aldose reductase LOC425137	XM_422928	1	0	high	0	1	0.000	2
aldose reductase LOC418171	XM_416402	1	2	0.500	0	1	0.000	4
CKLF-like MARVEL transmembrane domain containing 7	NM_001007894	2	0	high	0	2	0.000	3
ATP synthase mitochondrial F	XM_414815	1	1	1.000	0	1	0.000	3
F-box protein 7 FBXO7	NM_001012537	3	1	3.000	2	1	2.000	5
pelota homolog Drosophila PELO	NM_001031592	2	1	2.000	1	1	1.000	4
calsyntenin-3 LOC418297	XM_416520	2	2	1.000	0	2	0.000	5
casein kinase 1 epsilon CSNK1E	NM_204377	1	1	1.000	0	1	0.000	3
glucose transporter type 3 CEF-GT3	NM_205511	3	0	high	1	2	0.500	4
LOC421997 LOC421997	XM_429905	3	1	3.000	2	1	2.000	5
RIKEN cDNA B230118H07 LOC42	XM_421092	3	2	1.500	0	3	0.000	6
muscleblind-like 3 Drosophila MBNL3	NM_001012573	1	0	high	0	1	0.000	2
phosphodiesterase 3B cGMP-inhibited PDE3B	NM_001031182	1	1	1.000	0	1	0.000	3
discoidin CUB and LCCL domain containi	NM_001030783	2	0	high	2	0	high	3

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
SMAD mothers against DPP homolog 2 Drosophila	NM_204561	1	2	0.500	1	0	high	4
ligatin LGTN	NM_001006322	1	2	0.500	0	1	0.000	4
optineurin OPTN	NM_204236	2	1	2.000	0	2	0.000	4
nuclear protein p30 LOC4176	XM_415846	1	2	0.500	0	1	0.000	4
nucleoporin 85kDa NUP85	NM_001006426	1	1	1.000	0	1	0.000	3
heat shock 70kDa protein 4-like HSPA4L	NM_001012576	1	1	1.000	0	1	0.000	3
DEAH Asp-Glu-Ala-His box p	XM_422834	1	1	1.000	1	0	high	3
WD repeat domain 21 LOC4232	XM_421173	1	0	high	0	1	0.000	2
TXNRD3 protein LOC416031	XM_414371	1	2	0.500	0	1	0.000	4
hypothetical protein LOC422	XM_420805	1	1	1.000	0	1	0.000	3
CDNA sequence BC024814 LOC425865	NM_001031350	2	2	1.000	2	0	high	5
elastase 2A ELA2A	NM_001032390	21	13	1.615	9	1	2 0.750	35
non imprinted in Prader-Willi/Angelman syndrome 2	NM_001030809	2	1	2.000	1	1	1.000	4
vacuolar protein sorting	XM_426918	2	0	high	0	2	0.000	3
pancreatitis-induced protein	XM_415423	1	0	high	0	1	0.000	2
CD47 antigen Rh-related antigen integrin-associa	NM_001004388	3	0	high	1	2	0.500	4
DEAD Asp-Glu-Ala-Asp box polypeptide 42 DDX42	NM_001030926	2	0	high	1	1	1.000	3
Zgc:63829 LOC421130	NM_001031022	2	0	high	0	2	0.000	3
cisplatin resistance-associated overexp	NM_001031530	2	2	1.000	2	0	high	5
Expressed sequence AA960436	XM_413994	1	1	1.000	1	0	high	3
Ribonuclease P protein subun	XM_421667	1	2	0.500	0	1	0.000	4
polymerase DNA-directed delta 3 accessory subu	NM_001006284	1	2	0.500	1	0	high	4
angiotensin II receptor-associated	XM_417646	4	2	2.000	2	2	1.000	7
Probable ribosome biogenesis	XM_416515	2	0	high	1	1	1.000	3
phosphoglucomutase 2 PGM2	NM_001031383	1	0	high	0	1	0.000	2
adducin 1 alpha isoform d	XM_420826	1	1	1.000	0	1	0.000	3
Expressed sequence AI605202 LOC422523	NM_001031135	1	1	1.000	0	1	0.000	3
enolase 1 alpha ENO1	NM_205120	9	1	2 0.750	2	7	0.286	22
KCCR13L LOC416625	NM_001030643	1	0	high	0	1	0.000	2
protein phosphatase 2Aregulatory subunit B' PR	NM_001031371	1	1	1.000	0	1	0.000	3
tumor necrosis factor superf	XM_419125	1	1	1.000	0	1	0.000	3
synaptosomal-associatedprotein 91kDa homolog mo	NM_001012951	1	0	high	0	1	0.000	2
immune associated nucleotide	XM_427237	2	1	2.000	1	1	1.000	4
SERTA domain containing2 SERTAD2	NM_001031037	1	0	high	1	0	high	2
ARP2 actin-related protein 2 homolog yeast ACTR	NM_205224	1	1	1.000	0	1	0.000	3
solute carrier family 9(sodium/hydrogen exchanger	NM_001039275	2	1	2.000	1	1	1.000	4
hypothetical protein FLJ3211	XM_422491	1	1	1.000	0	1	0.000	3
RIKEN cDNA 0610037P05 LOC416598	NM_001006170	1	1	1.000	1	0	high	3
mitochondrial ribosomal prot	XM_420932	2	4	0.500	1	1	1.000	7
chromosome 3open reading frame 6 long	NM_001031315	2	0	high	1	1	1.000	3
growth factor receptor-bound protein 2 GRB2	NM_204411	2	1	2.000	1	1	1.000	4
LOC422206 LOC422206	XM_429926	1	0	high	1	0	high	2
ATP-binding cassette sub-family	XM_414701	2	0	high	1	1	1.000	3
tropomyosin 3 TPM3	NM_205446	1	0	high	0	1	0.000	2
A kinase PRKA anchor protein yotiao 9 AKAP9	NM_207179	1	1	1.000	1	0	high	3
kinesin-related microtuble-b	XM_417608	1	2	0.500	0	1	0.000	4
uncharacterized hypothalamus protein	NM_001006394	1	1	1.000	1	0	high	3
Leucine-zipper-like transcri	XM_419246	1	0	high	0	1	0.000	2
Aquaporin-1 AQP1	NM_001039453	1	2	0.500	0	1	0.000	4
neurochondrin NCDN	NM_001030901	1	0	high	0	1	0.000	2
sulfotransferase 1C SULT1C	NM_204601	1	0	high	0	1	0.000	2
DNA segment Chr 10 Johns	XM_417006	2	0	high	0	2	0.000	3
cofactor required for Sp1 transcriptional	NM_001006280	1	2	0.500	0	1	0.000	4
RAB9B member RAS oncogene f	XM_420182	1	2	0.500	0	1	0.000	4
RIKEN cDNA 1600014C10 LOC41	XM_414121	1	0	high	0	1	0.000	2
DGCR6 homolog DGCR6	NM_205040	1	1	1.000	0	1	0.000	3
FLJ20259 protein LOC416057	NM_001012782	1	2	0.500	0	1	0.000	4
cAMP responsive element bind	XM_424137	1	0	high	0	1	0.000	2
T-cell surface glycoprotein	XM_416858	3	3	1.000	0	3	0.000	7
IMMUNE-RESPONSIVE PROTEIN 1 LOC418812	NM_001030821	1	1	1.000	1	0	high	3
Niemann-Pick disease type C2 NPC2	NM_001031203	1	0	high	0	1	0.000	2
WD repeat domain 5 WDR5	NM_001006198	1	1	1.000	0	1	0.000	3

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
hypothetical protein FLJ2062	XM_419675	4	1	4.000	1	3	0.333	6
tafazzin LOC425281 partia	XM_423062	1	0	high	0	1	0.000	2
peptidylprolyl isomerase B cyclophilin B PPIB	NM_205461	3	1	3.000	1	2	0.500	5
ribosomal protein L3 RPL3	NM_001006241	4	5	0.800	1	3	0.333	10
ribosomal protein S6 RPS6	NM_205225	23	4	5.750	10	1	3 0.769	28
hypotheticalprotein FLJ11200 LOC42254	NM_001031136	1	0	high	0	1	0.000	2
hypothetical protein FLJ1335	XM_420703	1	1	1.000	0	1	0.000	3
NADH-cytochrome b5 reductase CYB5R	XM_420957	4	1	4.000	0	4	0.000	6
Rnps1 protein LOC416756	XM_415051	2	0	high	0	2	0.000	3
chromosome 9 open reading frame	XM_415471	3	2	1.500	1	2	0.500	6
Fas-interacting serine/threo	XM_419634	1	1	1.000	0	1	0.000	3
solute carrier family 7(cationic amino acid trans	NM_001030579	1	2	0.500	1	0	high	4
hypotheticalprotein FLJ38973 LOC42406	NM_001039306	1	0	high	0	1	0.000	2
hypotheticalprotein FLJ13193 LOC42712	NM_001006575	2	4	0.500	1	1	1.000	7
ceramide kinase CERK	NM_001031340	2	0	high	0	2	0.000	3
Hypotheticalprotein SB153 isoform 1	NM_001030627	1	0	high	0	1	0.000	2
probox protein PROBOX	NM_205252	1	0	high	0	1	0.000	2
carboxyl ester lipase bile salt-stimulated lipase	NM_001012997	3	5	0.600	0	3	0.000	9
leucine rich repeat containing 40 LRRC40	NM_001031295	1	2	0.500	0	1	0.000	4
DEAD Asp-Glu-Ala-Asp box polypeptide 5 DDX5	NM_204827	5	8	0.625	0	5	0.000	14
Nucleoporin Nup37 p37 LOC	XM_416326	1	1	1.000	1	0	high	3
ADP-ribosylation factor-like 6 interacting protein	NM_001006171	1	1	1.000	0	1	0.000	3
Kruppel-like factor 11 KLF11	NM_001006417	1	0	high	0	1	0.000	2
ring finger protein 126(RNF126	NM_001006338	2	0	high	0	2	0.000	3
chromosome 7open reading frame 28B; CG	NM_001006158	1	2	0.500	1	0	high	4
adenosine deaminase 1 ADAT1	NM_001012779	1	0	high	0	1	0.000	2
Peroxisomal acyl-coenzyme A	XM_417475	1	0	high	0	1	0.000	2
hypotheticalprotein FLJ10597 LOC42458	NM_001012599	1	1	1.000	1	0	high	3
Motile sperm domain containi	XM_420226	1	1	1.000	0	1	0.000	3
dystrobrevin binding protein 1 DTNBP1	NM_001006372	6	2	3.000	3	3	1.000	9
Plasma protease C1 inhibitor	XM_421063	1	0	high	0	1	0.000	2
monoamine oxidase A MAOA	NM_001030799	1	1	1.000	0	1	0.000	3
ClpX caseinolytic peptidase X homolog E. coli C	NM_001030552	2	4	0.500	0	2	0.000	7
2310047H23Rik protein LOC41	XM_414627	1	1	1.000	0	1	0.000	3
N-acetylneuraminate pyruvate lyase dihydrodipicol	NM_001031560	1	0	high	0	1	0.000	2
Protein phosphatase 1 regul	XM_421392	3	1	3.000	0	3	0.000	5
glutathione S-transferase A1 GSTA1	NM_001001777	1	1	1.000	1	0	high	3
Sentrin-specific protease 8	XM_413710	2	1	2.000	0	2	0.000	4
asrij LOC422764	XM_420717	1	0	high	0	1	0.000	2
mitogen-activated protein kinase kinase kinase 7	NM_001006240	1	0	high	0	1	0.000	2
TAP binding protein tapasin TAPBP	NM_001034816	6	8	0.750	2	4	0.500	15
guanine nucleotide exchange	XM_418283	1	0	high	1	0	high	2
serine/threonine kinase17a apoptosis-inducing	NM_001030995	1	0	high	1	0	high	2
ribosomal protein L37 LOC42	XM_424773	4	2	2.000	2	2	1.000	7
hypotheticalgene supported by CR386115	XM_429760	2	1	2.000	0	2	0.000	4
chromosome 6 open reading frame	XM_419737	1	1	1.000	0	1	0.000	3
protein tyrosine phosphatase receptor type C PT	NM_204417	1	2	0.500	1	0	high	4
ubiquitin-conjugating enzyme E2I UBC9 homolog	NM_204265	1	0	high	1	0	high	2
EPH receptor A3 EPHA3	NM_205430	1	0	high	1	0	high	2
Mitogen-activated protein kinase	XM_423647	1	0	high	0	1	0.000	2
calcium activated nucleotidase 1 CANT1	NM_001031581	1	0	high	0	1	0.000	2
solute carrier family 25 member 32 SLC25A32	NM_001031506	1	0	high	0	1	0.000	2
Yip1 domain family member 4 YIPF4	NM_001031058	1	2	0.500	0	1	0.000	4
prolyl endopeptidase PREP	NM_001006410	1	0	high	0	1	0.000	2
transcription factor p38 interacting	NM_001006275	2	0	high	0	2	0.000	3
Isovaleryl-CoA dehydrogenase	XM_420942	1	2	0.500	0	1	0.000	4
PEF protein with a long N-terminal	XM_417792	5	0	high	3	2	1.500	6
solute carrier family 25 mitochondrial carrier	NM_204231	6	4	1.500	1	5	0.200	11
Rho guanine nucleotide exchange factor GEF 3 AR	NM_001030595	1	1	1.000	1	0	high	3
haloacid dehalogenase-like hydrolase domain contai	NM_001031059	1	0	high	0	1	0.000	2
Mtap7 protein LOC422242	XM_420230	1	1	1.000	0	1	0.000	3
mitochondrial ribosomal protein	XM_417239	2	0	high	2	0	high	3
fetal globin inducing factor	XM_417191	1	2	0.500	1	0	high	4
XPMC2 prevents mitotic catas	XM_415436	2	2	1.000	1	1	1.000	5

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
riboflavin-binding protein LOC396449	NM_205463	3	0	high	1	2	0.500	4
golgi reassembly stacking protein 1 65kDa GORASP	NM_001030963	1	0	high	1	0	high	2
spindle assembly 6 homolog C. elegans SASS6	NM_001031273	1	0	high	1	0	high	2
chromosome 10 open reading frame	XM_424678	2	0	high	2	0	high	3
capping protein actin filament muscle Z-line	NM_205437	1	0	high	0	1	0.000	2
40S ribosomal protein S2 LO	XM_414845	11	5	2.200	3	8	0.375	17
methylcrotonoyl-Coenzyme A carboxylase	NM_001031565	1	1	1.000	1	0	high	3
Uroporphyrinogen decarboxyla	XM_422430	1	1	1.000	0	1	0.000	3
IQ calmodulin-binding motif	XM_422091	3	2	1.500	1	2	0.500	6
coproporphyrinogen oxidase	XM_416596	1	1	1.000	0	1	0.000	3
Down syndrome critical region gene 2 DSCR2	NM_001012543	2	3	0.667	0	2	0.000	6
hypotheticalgene supported by CR353050	XM_429928	2	0	high	0	2	0.000	3
TRAF-type zinc finger domain containing 1 TRAFD1	NM_001006191	1	0	high	0	1	0.000	2
CD83 antigen precursor Cell	XM_418929	1	0	high	0	1	0.000	2
CUG triplet repeat binding protein 2 CUGBP2	NM_204260	4	1	4.000	3	1	3.000	6
hypothetical protein FLJ2237	XM_418487	1	1	1.000	0	1	0.000	3
Phospholipase A2 precursor	XM_415272	1	0	high	1	0	high	2
ubiquinol--cytochrome c	XM_414356	4	2	2.000	3	1	3.000	7
ribonuclease T2	NM_001039491	1	2	0.500	0	1	0.000	4
dual specificity phosphatase	XM_423122	1	1	1.000	0	1	0.000	3
hypotheticalprotein MGC13096 LOC41578	NM_001030572	4	1	4.000	1	3	0.333	6
chromosome 6open reading frame 106 iso	NM_001030919	2	3	0.667	1	1	1.000	6
superoxide dismutase 2mitochondrial SOD2	NM_204211	5	5	1.000	1	4	0.250	11
MAP kinase-activated protein	XM_417976	1	2	0.500	1	0	high	4
thiopurine methyltransferase	XM_418921	1	0	high	0	1	0.000	2
hypothetical protein LOC420	XM_418965	3	1	3.000	1	2	0.500	5
DKFZp434C0328 protein LOC41	XM_416574	11	10	1.100	5	6	0.833	22
ankyrin repeat and MYNDdomain containing 2 ANKMY	NM_001030979	1	2	0.500	0	1	0.000	4
coagulation factor II thrombin receptor-like 1	NM_001012608	4	1	4.000	2	2	1.000	6
complement component 1 q	XM_425756	1	2	0.500	1	0	high	4
Phosphopantothenate--cystein	XM_417660	1	1	1.000	0	1	0.000	3
craniofacial development protein 1 CFDP1	NM_001001189	1	0	high	0	1	0.000	2
KIAA0117 protein LOC426311	XM_423974	1	1	1.000	0	1	0.000	3
soc-2 suppressor of clear homolog C. elegans SH	NM_001031236	5	1	5.000	0	5	0.000	7
Expressed sequence AI314976 LOC419916	NM_001030921	2	3	0.667	2	0	high	6
hypothetical protein MGC3212	XM_415743	1	1	1.000	0	1	0.000	3
mitochondrial ribosomalprotein L3 MRPL3	NM_001006366	2	0	high	0	2	0.000	3
heat shock 70kDa protein 9B mortalin-2 HSPA9B	NM_001006147	6	3	2.000	2	4	0.500	10
hypotheticalgene supported by BX930473	XM_429769	1	1	1.000	0	1	0.000	3
Hypothetical protein MGC5625	XM_421463	1	0	high	0	1	0.000	2
LOC424434 LOC424434	XM_422276	2	1	2.000	2	0	high	4
IMP inosine monophosphate dehydrogenase 2 IMPDH	NM_001030601	1	1	1.000	0	1	0.000	3
Cytochrome B-245 heavy chain	XM_416783	1	0	high	1	0	high	2
RIKEN cDNA 1110032A13 LOC420806	NM_001012861	2	0	high	0	2	0.000	3
zinc binding alcohol dehydro	XM_419096	1	0	high	0	1	0.000	2
BM-011 protein LOC425020	XM_422823	1	2	0.500	1	0	high	4
topoisomerase DNA I mitochondrial TOP1MT	NM_001001300	2	2	1.000	1	1	1.000	5
LOC398726 protein LOC417984	XM_416223	1	1	1.000	0	1	0.000	3
ankyrin repeat and SOCSbox-containing 9 ASB9	NM_001006262	2	0	high	1	1	1.000	3
BC282485_1 LOC424880	NM_001006546	2	3	0.667	2	0	high	6
Rpl10a-prov protein LOC4198	XM_418020	7	6	1.167	4	3	1.333	14
hypothetical protein LOC2708	XM_416063	2	1	2.000	2	0	high	4
RIKEN cDNA 2610318G18 LOC42	XM_419233	1	1	1.000	0	1	0.000	3
histone 2 H2ac LOC417955	XM_416195	1	1	1.000	1	0	high	3
RIKEN cDNA 5830418K08 gene	XM_417197	5	3	1.667	3	2	1.500	9
complement subcomponent C1Q	XM_417654	1	0	high	0	1	0.000	2
proliferation-associated	XM_423059	3	0	high	0	3	0.000	4
cytochrome P450 2D20 LOC417	XM_416219	2	3	0.667	0	2	0.000	6
leukocyte cell-derived chemotaxin 2 LECT2	NM_205478	1	1	1.000	0	1	0.000	3
discs large homolog 7	XM_421446	1	1	1.000	1	0	high	3
hypotheticalgene supported by CR387752	XM_429513	1	0	high	1	0	high	2

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
coilin; coilin p80 LOC41740	XM_415654	5	6	0.833	1	4	0.250	12
EF hand domain containing 2	XM_428222	1	0	high	0	1	0.000	2
ATP synthase subunit B LOC4	XM_417993	4	4	1.000	0	4	0.000	9
hypothetical protein BC00492	XM_418645	2	0	high	1	1	1.000	3
hypothetical protein LOC419	XM_417659	1	0	high	0	1	0.000	2
thymopoietin TMPO	NM_001006235	1	1	1.000	0	1	0.000	3
28S ribosomal protein S31	XM_417081	2	4	0.500	0	2	0.000	7
RAP2C member of RAS oncogene family RAP2C	NM_001012572	2	0	high	1	1	1.000	3
adult alpha D globin LOC416651	NM_001004375	6	3	2.000	1	5	0.200	10
MTERF domain containing1 MTERFD1	NM_001006348	1	1	1.000	0	1	0.000	3
apolipoprotein D APOD	NM_001011692	1	0	high	1	0	high	2
isocitrate dehydrogenase 2 NADP+	NM_001031599	2	1	2.000	0	2	0.000	4
hypothetical protein 9530023	XM_415188	3	1	3.000	1	2	0.500	5
aspartylglucosaminidase(AGA	NM_001006445	2	2	1.000	1	1	1.000	5
MTA3 protein LOC421395	XM_419452	1	1	1.000	1	0	high	3
Ammd protein LOC422088	XM_420090	2	0	high	0	2	0.000	3
ovoglycoprotein OGCHI	NM_204541	1	1	1.000	0	1	0.000	3
GATA zinc finger domaincontaining 2A GATAD2A	NM_001012552	2	0	high	1	1	1.000	3
denticleless homolog Drosophila DTL	NM_001031048	1	0	high	0	1	0.000	2
novel amplified in breast ca	XM_417504	1	0	high	1	0	high	2
proopiomelanocortin adrenocorticotropin/ beta-lip	NM_001031098	4	5	0.800	1	3	0.333	10
tumor necrosis factor receptor superfamily member	NM_204439	1	0	high	0	1	0.000	2
acyl-Coenzyme A dehydrogenase family member 9 AC	NM_001006136	2	3	0.667	0	2	0.000	6
ubiquitin-conjugating enzyme	XM_421525	1	0	high	0	1	0.000	2
protein phosphatase 1B formerly 2C magnesium	NM_001031052	1	1	1.000	0	1	0.000	3
leprecan-like 1 LEPREL1	NM_001001530	1	0	high	1	0	high	2
DNA-directed polymerases III	XM_415021	1	1	1.000	0	1	0.000	3
prostate stem cell antigen	XM_418414	2	0	high	0	2	0.000	3
Transcription intermediary	XM_416340	1	1	1.000	0	1	0.000	3
transforming acidic coiled-coil containing protei	NM_001004429	3	4	0.750	1	2	0.500	8
nephronectin short isoform	XM_420498	1	1	1.000	0	1	0.000	3
UDP-glucuronate decarboxylas	XM_416926	2	1	2.000	0	2	0.000	4
zinc finger CCCH type domain	XM_416342	1	0	high	0	1	0.000	2
hypothetical protein DKFZp54	XM_417165	1	1	1.000	0	1	0.000	3
ATP/GTP-binding protein(AI462438	NM_001012292	1	2	0.500	1	0	high	4
sulfatase modifying factor 2	XM_415788	4	2	2.000	1	3	0.333	7
ribosomal protein S3 RPS3	NM_001030836	9	8	1.125	2	7	0.286	18
MRE11 meiotic recombination 11 homolog A S. cerev	NM_204778	1	1	1.000	1	0	high	3
NADH dehydrogenase ubiquinone Fe-S protein 1 75	NM_001006518	6	7	0.857	2	4	0.500	14
mitochondrial carrier homolog 2 C. elegans MTCH	NM_204808	3	1	3.000	2	1	2.000	5
brother of CDO LOC418361	XM_416581	1	1	1.000	0	1	0.000	3
uroplakin 3B isoform b; urop	XM_415762	3	4	0.750	2	1	2.000	8
hypotheticalprotein BC009331 LOC42290	NM_001031168	1	0	high	0	1	0.000	2
DnaJ Hsp40 homolog	XM_417034	1	1	1.000	1	0	high	3
aldo-keto reductase AKR	NM_204629	1	0	high	0	1	0.000	2
MGC68903 protein LOC420644	XM_418743	1	1	1.000	1	0	high	3
Tetratricopeptide repeat domain	XM_414484	1	1	1.000	0	1	0.000	3
retinoblastoma binding protein 4 RBBP4	NM_204852	2	1	2.000	0	2	0.000	4
C-terminal binding protein 2	XM_421817	1	1	1.000	0	1	0.000	3
glutamate receptor ionotrop	XM_413788	1	2	0.500	1	0	high	4
mitochondrial ribosomal prot	XM_418869	1	0	high	0	1	0.000	2
Hypothetical protein D10Ertd	XM_419766	1	2	0.500	0	1	0.000	4
nucleoporin 98kD isoform 1	XM_428171	1	2	0.500	0	1	0.000	4
Tripartite motif protein 3	XM_422922	1	0	high	1	0	high	2
RIKEN cDNA G630055P03 gene	XM_414020	1	0	high	1	0	high	2
RIKEN cDNA 1200011I18 LOC418842	NM_001006270	1	1	1.000	0	1	0.000	3
phytanoyl-CoA hydroxylase protein	XM_424238	1	0	high	1	0	high	2
t-complex 1 TCP1	NM_001006405	2	3	0.667	1	1	1.000	6
crystallin beta B2 CRYBB2	NM_205175	1	2	0.500	0	1	0.000	4
Transcript increased in	XM_420103	1	1	1.000	0	1	0.000	3
Krueppel-like factor 5	XM_417013	1	0	high	1	0	high	2

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
CGI-105 protein LOC426684	XM_424309	1	0	high	0	1	0.000	2
NADP-dependent malic enzyme	XM_417212	1	1	1.000	1	0	high	3
small acidic protein LOC395520	NM_204758	4	3	1.333	2	2	1.000	8
phosphoserine phosphatase	XM_415786	2	0	high	0	2	0.000	3
LOC418260 LOC418260	XM_429563	3	0	high	3	0	high	4
calbindin 1 28kDa CALB1	NM_205513	1	0	high	0	1	0.000	2
Splicing factor arginine	XM_417951	1	1	1.000	1	0	high	3
Hexaprenyldihydroxybenzoate	XM_419824	1	1	1.000	0	1	0.000	3
complement component 1s subcomponent C1S	NM_001030777	1	0	high	0	1	0.000	2
LOC418413 LOC418413	NM_001006255	1	1	1.000	0	1	0.000	3
protein phosphatase 1 regulatory inhibitor subu	NM_205123	1	0	high	0	1	0.000	2
nuclear receptor coactivator 4 NCOA4	NM_001006495	1	2	0.500	0	1	0.000	4
phosphatidylinositol glycan class K PIGK	NM_001031278	1	0	high	1	0	high	2
p37NB protein P37NB	XM_415965	1	2	0.500	0	1	0.000	4
synthetase-like beta subunit FARSLB	NM_001006543	1	2	0.500	0	1	0.000	4
YRDC protein LOC419610	XM_417757	1	2	0.500	0	1	0.000	4
tumor necrosis factor alpha-induced protein 8-lik	NM_001006343	1	0	high	0	1	0.000	2
hypothetical protein FLJ2356	XM_416107	6	1	6.000	5	1	5.000	8
THAP domain containing 9 LO	XM_420555	1	0	high	0	1	0.000	2
glycoprotein 55 LOC415316	XM_413703	1	0	high	1	0	high	2
tyrosine 3-monooxygenase/tryptophan 5-monooxygenas	NM_001006219	6	1	6.000	3	3	1.000	8
synaptosomal-associatedprotein 29kDa SNAP29	NM_001030652	1	0	high	1	0	high	2
splicing factor arginine/serine-rich 6 SFRS6	NM_001030843	1	0	high	0	1	0.000	2
hypothetical protein MDS025	XM_417216	1	1	1.000	1	0	high	3
replication factor C activator 1 3 38kDa RFC3	NM_001006276	1	0	high	1	0	high	2
Opa-interacting protein 5	XM_421136	1	1	1.000	0	1	0.000	3
PAK1 interacting protein 1 PAK1IP1	NM_001030999	2	1	2.000	0	2	0.000	4
mal T-cell differentiation protein 2 MAL2	NM_001012555	1	0	high	0	1	0.000	2
NAD(P dependent steroid	XM_420279	1	2	0.500	0	1	0.000	4
glyoxylase 1; glyoxalase 1	XM_419481	1	2	0.500	1	0	high	4
GATA binding protein 5 GATA5	NM_205421	1	0	high	0	1	0.000	2
N-acetylgalactosaminyltransf	XM_418541	1	1	1.000	0	1	0.000	3
Ubl carboxyl-terminal hydrol	XM_416398	1	2	0.500	0	1	0.000	4
ribonuclease H1	NM_204998	1	1	1.000	0	1	0.000	3
lanosterol synthase 23-oxidosqualene-lanosterol	NM_001006514	1	0	high	1	0	high	2
PRO1853 protein isoform 1 L	XM_419525	2	2	1.000	1	1	1.000	5
outer dense fiber of sperm	XM_418368	2	2	1.000	0	2	0.000	5
splicing endonuclease 2homolog SEN2 S. cerevisi	NM_001030594	3	1	3.000	0	3	0.000	5
retinol binding protein 7	XM_417606	3	1	3.000	2	1	2.000	5
Secernin 1 LOC420635	XM_418734	1	1	1.000	1	0	high	3
CLK-1 LOC416609	XM_414911	1	1	1.000	1	0	high	3
mature avidin LOC396260	NM_205320	4	3	1.333	2	2	1.000	8
annexin A2 ANXA2	NM_205351	2	2	1.000	2	0	high	5
eukaryotic translation elongation factor 2 EEF2	NM_205368	2	0	high	1	1	1.000	3
endothelial-derived gene 1	XM_423123	2	1	2.000	0	2	0.000	4
tec protein tyrosine kinase TEC	NM_001030372	1	1	1.000	0	1	0.000	3
hypotheticalprotein LOC421135	NM_001031023	1	0	high	1	0	high	2
actin alpha cardiac; alpha	XM_421217	1	1	1.000	1	0	high	3
lymphoid transcription factor AIOLOS	NM_204820	1	0	high	1	0	high	2
divalent cation tolerant	XM_415407	1	1	1.000	1	0	high	3
ATP citrate lyase ACLY	NM_001030540	1	1	1.000	0	1	0.000	3
RIKEN cDNA 1110012E06 LOC419466	NM_001030881	1	0	high	0	1	0.000	2
ovoinhibitorprecursor [validated]	NM_001030612	2	3	0.667	1	1	1.000	6
Secretory granule proteoglyc	XM_421576	2	2	1.000	0	2	0.000	5
selenoprotein K SELK	NM_001025441	1	0	high	1	0	high	2
guanylate binding protein GBP	NM_204652	3	1	3.000	0	3	0.000	5
GATS protein LOC417483	XM_415731	1	1	1.000	1	0	high	3
cytochrome c oxidase subunit	XM_429181	1	1	1.000	0	1	0.000	3
down-regulated in metastasis	XM_416174	1	0	high	0	1	0.000	2
cartilage associated protein CRTAP	NM_205100	1	1	1.000	0	1	0.000	3
guanine nucleotide binding protein G protein	NM_001012793	1	1	1.000	1	0	high	3
putative N-acetyltransferase	XM_427317	1	0	high	0	1	0.000	2
hypothetical protein LOC416	XM_414751	1	1	1.000	0	1	0.000	3

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
interferon regulatory factor 4 IRF4	NM_204299	2	3	0.667	1	1	1.000	6
succinate-CoA ligase GDP-forming alpha subunit	NM_001012892	1	2	0.500	0	1	0.000	4
LOC420068 LOC420068	XM_429732	2	1	2.000	0	2	0.000	4
ring finger protein 128	XM_420351	1	1	1.000	0	1	0.000	3
nin one binding protein	XM_414227	1	0	high	0	1	0.000	2
lysophosphatidic acid	XM_416667	2	2	1.000	0	2	0.000	5
mitochondrial ribosomal protein	XM_424803	1	0	high	0	1	0.000	2
Transcription factor BTF3	XM_423823	2	0	high	1	1	1.000	3
interleukin-7 receptor precursor	XM_423732	1	0	high	1	0	high	2
cytidine deaminase LOC41745	XM_415706	3	0	high	0	3	0.000	4
Sorting nexin 6 TRAF4-associated	XM_421235	1	1	1.000	0	1	0.000	3
E3 protein LOC419552	XM_417701	1	1	1.000	1	0	high	3
Synaptotagmin XI SytXI LO	XM_426721	1	1	1.000	0	1	0.000	3
2-amino-3-ketobutyrate coenzyme	XM_425478	2	3	0.667	1	1	1.000	6
4833424P18Rik protein LOC42	XM_422785	7	3	2.333	1	6	0.167	11
NADH dehydrogenase ubiquinone 1 alpha subcomplex	NM_001031247	1	1	1.000	1	0	high	3
cyclin C LOC421791	XM_419818	2	0	high	1	1	1.000	3
mitochondrial ribosomal prot	XM_419327	2	1	2.000	1	1	1.000	4
RIKEN cDNA 1200009B18	XM_415941	1	1	1.000	0	1	0.000	3
vaccinia related kinase1 VRK1	NM_001006485	1	0	high	0	1	0.000	2
thioredoxin domain containing 5 TXNDC5	NM_001006374	2	1	2.000	0	2	0.000	4
sperm specific antigen 2	XM_421971	1	2	0.500	0	1	0.000	4
phosphonoformate immuno-associated prot	NM_001012828	1	1	1.000	1	0	high	3
DEAD box polypeptide 17 isof	XM_416260	1	1	1.000	1	0	high	3
FLJ22353 LOC424581	XM_422420	1	2	0.500	0	1	0.000	4
RIKEN cDNA 1300013J15 LOC417617	NM_001030720	1	0	high	1	0	high	2
NADH2 dehydrogenase ubiquin	XM_424129	1	0	high	1	0	high	2
KIAA0882 protein LOC422450	XM_420416	2	2	1.000	0	2	0.000	5
maleylacetoacetate isomerase	XM_421288	1	0	high	0	1	0.000	2
PLCPI=cysteine proteinase	XM_416493	2	1	2.000	0	2	0.000	4
estrogen receptor binding	XM_422333	2	1	2.000	0	2	0.000	4
hypothetical protein dJ122O8	XM_419836	1	0	high	0	1	0.000	2
KIAA1705 protein LOC423591	XM_421479	1	0	high	1	0	high	2
hypothetical protein RP4-622	XM_417801	3	3	1.000	2	1	2.000	7
RIKEN cDNA 2610528E23 LOC41	XM_416600	1	0	high	1	0	high	2
myosin light polypeptide 4 alkali; atrial embry	NM_205479	2	0	high	1	1	1.000	3
ribosomal protein L4 RPL4	NM_001007479	8	7	1.143	5	3	1.667	16
calmodulin 2 phosphorylase kinase delta CALM2	NM_205005	3	2	1.500	1	2	0.500	6
solute carrier family 25 member 13 citrin SLC2	NM_001012949	1	1	1.000	0	1	0.000	3
COMM domain containing 8	XM_420723	1	0	high	0	1	0.000	2
cell division cycle associated 1 CDCA1	NM_204478	1	0	high	0	1	0.000	2
origin recognition complex	XM_414114	1	2	0.500	0	1	0.000	4
GLI pathogenesis-related 1 glioma GLIPR1	NM_001030743	4	1	4.000	2	2	1.000	6
FLJ00156 protein LOC423781	XM_421653	1	0	high	0	1	0.000	2
Hypothetical protein MGC6373	XM_421719	1	0	high	1	0	high	2
NADH dehydrogenase LOC41639	XM_414705	1	0	high	1	0	high	2
ribosomal protein large P1 RPLP1	NM_205322	1	2	0.500	0	1	0.000	4
proteasome prosome macropain activator subunit	NM_001012550	2	0	high	1	1	1.000	3
4-hydroxyphenylpyruvate	XM_415144	1	1	1.000	0	1	0.000	3
Mphase phosphoprotein 6	XM_414172	1	0	high	0	1	0.000	2
DLNB14 LOC419787	XM_417924	1	0	high	0	1	0.000	2
AT rich interactive domain 5B MRF1-like ARID5B	NM_001031220	3	2	1.500	1	2	0.500	6
hypotheticalgene supported by CR391196	XM_429449	2	1	2.000	0	2	0.000	4
UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-a	NM_001031578	2	1	2.000	2	0	high	4
Ras-related protein Rab-7	XM_414359	2	1	2.000	1	1	1.000	4
hypothetical protein MGC3526	XM_420180	1	2	0.500	0	1	0.000	4
RIO kinase 2 yeast RIOK2	NM_001006581	1	0	high	0	1	0.000	2
small nuclear ribonucleoprotein polypeptides B and	NM_204599	4	2	2.000	1	3	0.333	7
26S proteasome non-ATPase	XM_420921	1	0	high	1	0	high	2
Gprotein gamma-5 subunit	XM_422375	1	1	1.000	1	0	high	3
uncharacterized hematopoieti	XM_420312	4	2	2.000	1	3	0.333	7
tissue inhibitor of metalloproteinase 2 TIMP2	NM_204298	1	1	1.000	1	0	high	3

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
chromosome 14 open reading	XM_421234	2	1	2.000	1	1	1.000	4
LOC422091	XM_420093	1	2	0.500	1	0	high	4
TIMP metallopeptidase inhibitor 3 Sorsby fundus	NM_205487	7	6	1.167	5	2	2.500	14
uridine-cytidine kinase1-like 1 UCKL1	NM_001037830	2	4	0.500	1	1	1.000	7
ADMP LOC425008	XM_422812	1	1	1.000	1	0	high	3
electron-transfer-flavoprotein alpha polypeptide	NM_001030543	1	0	high	0	1	0.000	2
BAG-family molecular chapero	XM_419051	1	0	high	0	1	0.000	2
FLJ14007 protein LOC420200	XM_418311	1	2	0.500	0	1	0.000	4
hypothetical protein LOC6392	XM_416240	1	0	high	0	1	0.000	2
CG1275-PB LOC423025	XM_420955	1	1	1.000	0	1	0.000	3
RIKEN cDNA 2410004L22 LOC42	XM_418254	1	1	1.000	0	1	0.000	3
Protein KIAA0586 LOC423539	XM_421432	1	2	0.500	1	0	high	4
connector enhancer of kinase suppressor	NM_001006434	3	2	1.500	2	1	2.000	6
hypothetical protein D10Ertd	XM_416121	1	2	0.500	0	1	0.000	4
hypotheticalprotein FLJ10656; cyclin-d	NM_001031005	2	2	1.000	0	2	0.000	5
CG10958-like LOC422009	XM_420016	1	0	high	1	0	high	2
G10 protein homolog EDG-2	XM_414798	1	2	0.500	0	1	0.000	4
densin-180-like protein LOC	XM_429138	1	0	high	1	0	high	2
60 kDa heat shock protein mitochondria	NM_001012916	5	4	1.250	1	4	0.250	10
cathepsin Y LOC419311	XM_417483	3	5	0.600	0	3	0.000	9
transmembrane emp24 protein transport domain conta	NM_001007956	2	0	high	2	0	high	3
KIAA1802 protein LOC418733	XM_416932	2	2	1.000	0	2	0.000	5
hypothetical protein FLJ1003	XM_413922	2	1	2.000	1	1	1.000	4
Heat shock protein 67B2 LOC421792	NM_001006411	1	0	high	1	0	high	2
CD82 antigen CD82	NM_001008470	1	0	high	0	1	0.000	2
Single-stranded DNA binding	XM_416355	2	2	1.000	0	2	0.000	5
small nuclear activating	XM_421416	4	1	4.000	1	3	0.333	6
protein phosphatase 4 regulatory subunit 2 PPP4R	NM_001006142	4	1	4.000	4	0	high	6
RIKEN cDNA 2010001O09 LOC42	XM_422503	5	0	high	1	4	0.250	6
Arl6ip2 protein LOC426447	XM_424092	1	1	1.000	1	0	high	3
mitochondrial ribosomal protein	XM_420854	2	2	1.000	0	2	0.000	5
T-cell immunoglobulin and mucin domain containing	NM_001006149	5	5	1.000	2	3	0.667	11
U2 small nuclear ribonucleoprotein	XM_419331	5	3	1.667	2	3	0.667	9
mast cell proteinase-1 LOC4	XM_423728	2	2	1.000	1	1	1.000	5
chemokine C-C motif ligand 20 CCL20	NM_204438	1	0	high	0	1	0.000	2
mitochondrial ribosomalprotein L50 MRPL50	NM_001006583	2	2	1.000	0	2	0.000	5
P1 subunit LOC419992	XM_418114	2	1	2.000	0	2	0.000	4
5-methylaminomethyl-2-thiouridylate methyltransfer	NM_001031351	1	1	1.000	0	1	0.000	3
methyltransferase like 2 METTL2	NM_001006329	1	1	1.000	0	1	0.000	3
inhibin beta A activin A activin AB alpha polyp	NM_205396	2	0	high	0	2	0.000	3
prion protein interacting protein	XM_422418	1	0	high	0	1	0.000	2
maternal embryonic leucine zipper kinase MELK	NM_001031509	1	1	1.000	0	1	0.000	3
pleckstrin homology domain containing family F	NM_001030947	2	0	high	1	1	1.000	3
receptor kinase LOC426000	NM_001031353	2	1	2.000	1	1	1.000	4
vacuolar protein sorting 35 yeast VPS35	NM_001005842	1	0	high	0	1	0.000	2
peroxisomal trans-2-enoyl-CoA reductase PECR	NM_001006522	1	0	high	1	0	high	2
BH3 interacting domain death agonist BID	NM_204552	3	3	1.000	1	2	0.500	7
Extracellular superoxide	XM_420760	2	1	2.000	0	2	0.000	4
cytidine deaminase CDD	NM_204933	1	2	0.500	0	1	0.000	4
Ras association RalGDS/AF-6 domain family 2 RAS	NM_001030884	4	2	2.000	3	1	3.000	7
dystrobrevin binding protein	XM_417359	1	1	1.000	1	0	high	3
myosin light polypeptide 3 alkali; ventricular	NM_205159	1	1	1.000	0	1	0.000	3
keratocan KERA	NM_204176	1	0	high	0	1	0.000	2
chromosome 7 open reading frame	XM_418776	1	0	high	0	1	0.000	2
chromosome 9open reading frame 76	NM_001031612	1	0	high	0	1	0.000	2
dendritic cell protein LOC421602	NM_001006406	2	3	0.667	0	2	0.000	6
KIAA1530 protein LOC422903	XM_420845	2	1	2.000	0	2	0.000	4
coatomer protein complex subunit alpha COPA	NM_001031405	1	2	0.500	0	1	0.000	4
SIGIRR LOC422995	XM_420927	1	0	high	0	1	0.000	2
PINCH-1 LOC418729	XM_416928	1	1	1.000	0	1	0.000	3
Secernin 2 LOC425759 partial	XM_423480	5	1	5.000	4	1	4.000	7

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
Hypothetical protein CBG0897	XM_414184	3	1	3.000	1	2	0.500	5
LOC417860 LOC417860	XM_429533	1	2	0.500	1	0	high	4
Tcell receptor alpha LOC41	XM_418060	1	0	high	1	0	high	2
Serine/threonine-protein kinase	XM_414252	3	0	high	0	3	0.000	4
REV1-like yeast REV1L	NM_001030811	1	1	1.000	0	1	0.000	3
cell division cycle 40 homolog yeast CDC40	NM_001006407	1	0	high	1	0	high	2
hypothetical protein FLJ2072	XM_422364	1	0	high	1	0	high	2
Ab2-416 LOC418398	NM_001006253	1	0	high	0	1	0.000	2
plasticity related gene 3;	XM_424888	1	1	1.000	0	1	0.000	3
hypothetical protein LOC425	XM_423139	1	0	high	0	1	0.000	2
Solute carrier family 12 member	XM_424716	2	1	2.000	2	0	high	4
ribophorin II RPN2	NM_001006288	2	2	1.000	1	1	1.000	5
calcium modulating ligand CAMLG	NM_204962	1	1	1.000	0	1	0.000	3
syndecan 2 heparan sulfate proteoglycan 1 cell	NM_001001462	1	1	1.000	0	1	0.000	3
MGC64570 protein LOC415776	XM_414138	5	0	high	3	2	1.500	6
MGC68821 protein LOC416477	XM_414782	1	0	high	0	1	0.000	2
LOC420990 LOC420990	XM_429802	1	0	high	1	0	high	2
Sh3yl1 LOC421908	XM_419926	1	1	1.000	1	0	high	3
adenylate kinase EC 2.7.4.3	XM_425786	1	0	high	0	1	0.000	2
signal transducer and activator of transcription 4	NM_001012914	1	0	high	0	1	0.000	2
amino acid feature	XM_416058	1	1	1.000	1	0	high	3
WD repeat and FYVE domain containing 1 WDFY1	NM_001031310	2	4	0.500	1	1	1.000	7
hypotheticalgene supported by CR385186	XM_429872	1	1	1.000	1	0	high	3
DAZ associated protein 1 LOC427266	NM_001031428	6	1	6.000	2	4	0.500	8
RIKEN cDNA 6330416G13 gene	XM_415524	2	0	high	0	2	0.000	3
ubiquitin specific peptidase 10 USP10	NM_001006130	3	2	1.500	1	2	0.500	6
RAB-interacting factor LOC4	XM_419250	1	0	high	0	1	0.000	2
ACN9 homolog LOC420572	XM_418673	2	2	1.000	1	1	1.000	5
synaptogyrin 3 SYNGR3	NM_001007834	3	1	3.000	2	1	2.000	5
Seizure 6-like protein precu	XM_415197	1	0	high	1	0	high	2
Phosphatidate cytidylyltrans	XM_417669	1	0	high	0	1	0.000	2
hypothetical protein MGC3508	XM_420702	7	1	7.000	0	7	0.000	9
pyrroline-5-carboxylate redu	XM_415641	2	2	1.000	0	2	0.000	5
LOC418128 LOC418128	XM_429551	1	1	1.000	0	1	0.000	3
organic solute transporter b	XM_413900	2	0	high	1	1	1.000	3
ralA binding protein 1 RALBP1	NM_001031575	1	0	high	0	1	0.000	2
RIKEN cDNA 1810012I05 LOC41	XM_417878	1	0	high	0	1	0.000	2
limb-bud andheart LOC421301	NM_001031038	2	0	high	1	1	1.000	3
Nucleoside diphosphate kinas	XM_416591	2	0	high	0	2	0.000	3
RIKEN cDNA 9530058B02 LOC42	XM_422971	1	1	1.000	0	1	0.000	3
5-aminoimidazole-4-carboxamide ribonucleotide form	NM_205178	3	3	1.000	1	2	0.500	7
asparagine synthetase ASNS	NM_001030977	2	1	2.000	1	1	1.000	4
hypotheticalgene supported by CR389209	XM_416629	1	0	high	0	1	0.000	2
heat shock protein 25 HSP25	NM_001010842	1	1	1.000	1	0	high	3
Probable G protein-coupled	XM_422842	1	0	high	0	1	0.000	2
Transforming growth factor-b	XM_418366	1	0	high	0	1	0.000	2
TYRO3 protein tyrosine kinase TYRO3	NM_204627	1	1	1.000	0	1	0.000	3
Nedd4-binding protein 3 N4BP3 LOC431	NM_001031607	1	0	high	0	1	0.000	2
pseudoautosomal GTP-binding	XM_416868	2	3	0.667	0	2	0.000	6
Ras-related protein Ral-B L	XM_422085	1	0	high	0	1	0.000	2
preimplantation protein3 PREI3	NM_001031557	1	2	0.500	1	0	high	4
RIKEN cDNA 1110046L09	XM_423808	1	0	high	0	1	0.000	2
mitochondrial ribosomal prot	XM_416895	1	0	high	0	1	0.000	2
solute carrier family 19 me	XM_422610	1	2	0.500	0	1	0.000	4
methyltransferase Cyt19	XM_421735	2	2	1.000	0	2	0.000	5
CGI-90 protein LOC420213	XM_418323	1	1	1.000	0	1	0.000	3
hyaluronan receptor - human	XM_414495	1	0	high	0	1	0.000	2
KIAA1630 protein LOC425149	XM_422940	1	0	high	0	1	0.000	2
hypotheticalgene supported by CR389756	XM_423439	2	0	high	1	1	1.000	3
endoplasmic reticulum chaper	XM_414514	1	0	high	0	1	0.000	2
complement component 3areceptor 1 C3AR1	NM_001030769	1	1	1.000	1	0	high	3
RIKEN cDNA 2010323F13 gene	XM_420935	1	1	1.000	0	1	0.000	3
stathmin-like 2 STMN2	NM_205181	1	1	1.000	0	1	0.000	3

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
Growth-arrest-specific prote	XM_414217	2	2	1.000	0	2	0.000	5
510-methenyltetrahydrofolat	XM_413857	2	2	1.000	0	2	0.000	5
Cytochrome c-type heme lyase CCHL	NM_001031275	1	2	0.500	0	1	0.000	4
Rap guanine nucleotide exchange factor	NM_001030982	1	2	0.500	1	0	high	4
transferrin TF	NM_205304	2	2	1.000	0	2	0.000	5
hypotheticalgene supported by CR389067	XM_429448	4	0	high	1	3	0.333	5
KIAA1105 protein LOC415998	XM_414340	1	0	high	1	0	high	2
hypothetical protein LOC428	XM_425814	1	0	high	0	1	0.000	2
Inner membrane protein OXA1L	XM_423587	4	1	4.000	1	3	0.333	6
biotinidase LOC420639	XM_418738	1	0	high	1	0	high	2
RIKEN cDNA 2010316F05 LOC42	XM_419287	2	4	0.500	0	2	0.000	7
3110023E09Rik protein LOC42	XM_421037	2	2	1.000	0	2	0.000	5
ARP5 actin-related protein 5 homolog yeast ACTR	NM_001008446	1	2	0.500	1	0	high	4
prosaposin variant Gaucher disease and variant me	NM_204811	2	3	0.667	0	2	0.000	6
lysozyme renal amyloidosis LYZ	NM_205281	7	3	2.333	4	3	1.333	11
1110063C11Rik protein LOC42	XM_424531	1	1	1.000	0	1	0.000	3
hypothetical protein DKFZp76	XM_423299	3	0	high	2	1	2.000	4
vascular cell adhesion molec	XM_422310	1	0	high	1	0	high	2
oxidative-stress responsive	XM_418527	1	2	0.500	0	1	0.000	4
AMP deaminase 1 Myoadenylat	XM_418010	1	0	high	1	0	high	2
KIAA1824 protein LOC428876	XM_426432	1	1	1.000	1	0	high	3
ureidopropionase beta LOC4	XM_415242	1	1	1.000	0	1	0.000	3
polymerase LOC423150	XM_421077	1	1	1.000	1	0	high	3
calcium-binding tyrosine pho	XM_419164	1	0	high	0	1	0.000	2
Hypothetical protein MGC7613	XM_417074	1	0	high	1	0	high	2
Hypothetical protein FLJ1315	XM_422341	1	1	1.000	0	1	0.000	3
nuclear fragile X mental retardation protein inter	NM_001030825	1	1	1.000	0	1	0.000	3
Aquaporin 7 Aquaporin-7 lik	XM_424498	1	2	0.500	0	1	0.000	4
hypothetical protein FLJ1329	XM_413987	1	0	high	1	0	high	2
adrenal hyoplasia protein DAX1 DAX1	NM_204593	4	1	4.000	1	3	0.333	6
Delta-like homolog LOC42345	XM_421369	2	3	0.667	0	2	0.000	6
FHA-HIT LOC395173	XM_429199	1	1	1.000	1	0	high	3
avidin LOC426220	XM_423883	5	0	high	5	0	high	6
NIF3L1 LOC424076	XM_421932	1	2	0.500	1	0	high	4
Ptcd2 protein LOC427217	XM_424804	1	2	0.500	0	1	0.000	4
solute carrier family 35 member B1 SLC35B1	NM_204514	3	2	1.500	1	2	0.500	6
midline 1 Opitz/BBB syndrome MID1	NM_204129	1	0	high	0	1	0.000	2
atrial natriuretic factor precursor LOC395765	NM_204925	2	2	1.000	0	2	0.000	5
lactate dehydrogenase B(LDHB	NM_204177	16	7	2.286	3	1	3 0.231	24
60S ribosomal protein L8 LO	XM_416772	32	10	2.300	9	1	4 0.643	34
lysosomal-associated membrane protein 2 LAMP2	NM_001001749	3	4	0.750	2	1	2.000	8
MGC69029 protein LOC417520	XM_415768	1	0	high	0	1	0.000	2
polymerase II DNA directed	XM_422761	1	1	1.000	1	0	high	3
hypothetical protein IMPACT	XM_419166	2	3	0.667	0	2	0.000	6
FK506 binding protein 4 59kDa FKBP4	NM_001006250	1	0	high	0	1	0.000	2
chromosome 10 open reading f	XM_421599	1	0	high	0	1	0.000	2
phosphoglycerate mutase EC 5.4.2.1 B	NM_001031556	4	2	2.000	0	4	0.000	7
hypotheticalprotein LOC422327	NM_001031124	1	0	high	1	0	high	2
Dipeptidyl-peptidase I	XM_417207	3	2	1.500	1	2	0.500	6
transcription factor GABP	XM_423392	1	2	0.500	0	1	0.000	4
arylsulfatase D ARSD	NM_204372	1	2	0.500	0	1	0.000	4
MANSC domain containing1 MANSC1	NM_001031381	4	2	2.000	1	3	0.333	7
RIKEN cDNA 4930430F08	XM_416129	1	1	1.000	1	0	high	3
dJ85M6.4 novel 58.3 KDA pro	XM_419630	2	0	high	2	0	high	3
solute carrier family 25 mitochondrial carrier	NM_001006443	1	0	high	1	0	high	2
DEAH Asp-Glu-Ala-His box polypeptide 15 DHX15	NM_001031159	1	0	high	0	1	0.000	2
processing of precursor 5	XM_415266	1	1	1.000	1	0	high	3
expressed sequence AW413431	XM_414221	3	3	1.000	0	3	0.000	7
hypothetical protein MGC9907	XM_421411	2	1	2.000	1	1	1.000	4
WNT1 inducible signaling pathway protein 1 WISP1	NM_001024579	1	0	high	1	0	high	2
CDNA sequence BC027073 LOC4	XM_422493	1	2	0.500	0	1	0.000	4
decay-accelerating factor G	XM_417981	3	1	3.000	1	2	0.500	5

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
hypotheticalgene supported by CR390114	XM_429784	1	1	1.000	0	1	0.000	3
hypothetical protein A430083	XM_414413	2	3	0.667	2	0	high	6
kinesin likeprotein LOC423809	NM_001031230	1	0	high	1	0	high	2
muscle specific ring finger	XM_424369	1	1	1.000	0	1	0.000	3
developmentally regulated GTP binding protein 2 D	NM_001030634	1	2	0.500	0	1	0.000	4
6720435I21Rik protein LOC41	XM_416366	1	1	1.000	0	1	0.000	3
Myosin light chain kinase	XM_425838	2	1	2.000	0	2	0.000	4
Mitochondrial 28S ribosomal	XM_424290	2	0	high	1	1	1.000	3
RNF121 protein LOC419092	XM_417284	2	0	high	1	1	1.000	3
S-adenosylhomocysteine hydrolase-like 1 AHCYL1	NM_001030913	1	1	1.000	0	1	0.000	3
TGFB-induced factor TALE family homeobox TGIF	NM_205379	1	1	1.000	0	1	0.000	3
CG9967-PA LOC415966	XM_414310	1	0	high	0	1	0.000	2
tectorin beta TECTB	NM_205363	1	0	high	0	1	0.000	2
NIMA never in mitosis gene a-related kinase 6 N	NM_001012531	1	1	1.000	0	1	0.000	3
60S ribosomal protein L17 L	XM_424454	4	4	1.000	1	3	0.333	9
prostaglandin-D synthase PGDS	NM_205011	1	1	1.000	0	1	0.000	3
hypotheticalgene supported by BX932049	XM_422395	1	1	1.000	0	1	0.000	3
Hypothetical protein MGC7575	XM_414294	1	0	high	1	0	high	2
Hep21 protein LOC395192	NM_204521	2	0	high	1	1	1.000	3
Sp3 transcription factor SP3	NM_204603	1	0	high	0	1	0.000	2
KIAA0582 protein LOC421274	XM_419343	2	2	1.000	1	1	1.000	5
PTK7 protein tyrosine kinase 7 PTK7	NM_001031035	3	0	high	1	2	0.500	4
THO complex subunit 2 Tho2	XM_420332	2	0	high	1	1	1.000	3
zinc finger FYVE domain con	XM_421128	5	1	5.000	1	4	0.250	7
small GTP binding protein RA	XM_419896	2	1	2.000	1	1	1.000	4
regulator of G-protein signalling 4 RGS4	NM_204385	2	0	high	0	2	0.000	3
Heat-shock protein beta-7 H	XM_427836	2	0	high	1	1	1.000	3
Putative GTP-binding protein	XM_416289	4	1	4.000	1	3	0.333	6
carboxy-terminal kinesin 2	XM_415326	4	2	2.000	0	4	0.000	7
RasGEF domain family member 1A RASGEF1A	NM_001031219	1	0	high	1	0	high	2
B-ATF LOC423364	XM_421279	1	2	0.500	1	0	high	4
CG12863-PA LOC422806	XM_420756	1	0	high	0	1	0.000	2
coagulation factor IX plasma thromboplastic compo	NM_204343	2	2	1.000	1	1	1.000	5
Myosin IXb Unconventional	XM_418252	2	0	high	0	2	0.000	3
Abl-219 LOC420538	XM_418640	2	2	1.000	0	2	0.000	5
RIKEN cDNA 2610015J01 LOC42	XM_421171	1	0	high	0	1	0.000	2
BC022687 protein LOC423498	XM_421401	2	1	2.000	2	0	high	4
replication factor C activator 1 1 145kDa RFC1	NM_001006456	3	2	1.500	2	1	2.000	6
transient receptor potential cation channel subfa	NM_001039317	2	0	high	0	2	0.000	3
laminin beta 2 laminin S LAMB2	NM_204166	1	1	1.000	1	0	high	3
ferritoid FTD	NM_204383	1	1	1.000	0	1	0.000	3
P25 protein LOC420800	XM_418894	1	0	high	0	1	0.000	2
procollagen type III N-endopeptidase PCOLN3	NM_001025440	2	2	1.000	0	2	0.000	5
caveolin 2 CAV2	NM_001007086	2	0	high	1	1	1.000	3
immune associated nucleotide	XM_418486	1	0	high	1	0	high	2
chromosome 1 open reading frame	XM_422229	5	1	5.000	3	2	1.500	7
potassium channel tetrameris	XM_414405	1	0	high	0	1	0.000	2
3-terminal phosphate cyclase	XM_424808	1	0	high	0	1	0.000	2
lipoprotein lipase LPL	NM_205282	1	1	1.000	0	1	0.000	3
KIAA1596 protein LOC423594	XM_421482	1	2	0.500	1	0	high	4
Ormdl2 protein LOC425059	NM_001031325	2	0	high	1	1	1.000	3
apolipoprotein AV; regenerat	XM_417939	6	0	high	4	2	2.000	7
transmembrane protein 66 TMEM66	NM_001031146	4	2	2.000	0	4	0.000	7
apo AI promoter B-region binding protein	XM_422765	1	0	high	0	1	0.000	2
Feather keratin I Keratin	XM_428216	1	0	high	0	1	0.000	2
Feline leukemia virus subgroup	XM_421280	1	0	high	0	1	0.000	2
FLJ44216 protein	XM_414554	2	3	0.667	1	1	1.000	6
glutathione peroxidase 4 phospholipid hydroperoxi	NM_204220	1	1	1.000	1	0	high	3
antizyme inhibitor 1 AZIN1	NM_001008729	1	0	high	1	0	high	2
hypothetical protein FLJ2034	XM_415842	1	0	high	1	0	high	2
DNA segment Chr 3 ERATO	XM_422361	1	0	high	0	1	0.000	2
Myotubularin related protein	XM_417799	1	1	1.000	0	1	0.000	3
proteasome prosome macropain 26S subunit ATPas	NM_001006494	4	4	1.000	2	2	1.000	9

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
vitronectin serum spreading factor somatomedin B	NM_205061	3	3	1.000	1	2	0.500	7
hypothetical protein MGC3040	XM_418792	2	0	high	0	2	0.000	3
Mannosidase beta A lysosom	XM_420666	1	1	1.000	0	1	0.000	3
hypothetical protein LOC418	XM_417166	1	0	high	1	0	high	2
RIKEN cDNA 2610003J06 LOC42	XM_424526	1	0	high	0	1	0.000	2
LOC418006 LOC418006	XM_429542	4	2	2.000	2	2	1.000	7
serine/threonine protein kin	XM_426558	2	0	high	0	2	0.000	3
v-yes-1 Yamaguchi sarcoma viral related oncogene	NM_001006390	2	0	high	1	1	1.000	3
hypotheticalgene supported by CR407377	XM_429342	1	0	high	0	1	0.000	2
centrosomal protein 2; centr	XM_417323	2	0	high	0	2	0.000	3
ribosomal protein L13 RPL13	NM_204999	8	8	1.000	4	4	1.000	17
axin 1 AXIN1	NM_204944	2	1	2.000	0	2	0.000	4
hypothetical protein FLJ3066	XM_418386	3	3	1.000	0	3	0.000	7
DNA segment Chr 8 ERATO Do	XM_420588	2	1	2.000	0	2	0.000	4
6-phosphogluconolactonase PGLS	NM_001031588	1	2	0.500	0	1	0.000	4
hypothetical protein FLJ2077	XM_420244	1	1	1.000	0	1	0.000	3
sema domain immunoglobulin domain Ig short	NM_204258	1	1	1.000	0	1	0.000	3
hypotheticalgene supported by CR386385	XM_430120	1	2	0.500	0	1	0.000	4
ring finger protein 4 RNF4	NM_001012889	1	0	high	0	1	0.000	2
Hypothetical protein KIAA013	XM_419580	2	1	2.000	0	2	0.000	4
Peripheral-type benzodiazepi	XM_423334	1	0	high	1	0	high	2
28S ribosomal protein S17	XM_415784	1	2	0.500	0	1	0.000	4
KIAA1093 protein LOC418010	XM_416246	1	0	high	1	0	high	2
lactate dehydrogenase A(LDHA	NM_205284	5	2	2.500	2	3	0.667	8
cytochrome b reductase 1	XM_421999	1	0	high	0	1	0.000	2
glutaredoxin 2 isoform 1	XM_422200	2	0	high	0	2	0.000	3
ENSANGP00000020885 LOC42491	XM_422728	3	0	high	0	3	0.000	4
cobl-related 1 LOC424182	XM_422028	1	2	0.500	0	1	0.000	4
Mic2l1 LOC422386	XM_420355	2	0	high	2	0	high	3
butyrylcholinesterase BCHE	NM_204646	1	0	high	0	1	0.000	2
mitochondrial ribosomal prot	XM_419623	3	1	3.000	1	2	0.500	5
Burkitt lymphoma receptor 1	XM_420151	1	0	high	1	0	high	2
DnaJ Hsp40 homolog subfam	XM_424624	2	0	high	1	1	1.000	3
DEAD Asp-Glu-Ala-As box polypeptide 19B DDX19B	NM_001006568	5	5	1.000	3	2	1.500	11
Friend of GATA LOC415837	XM_414198	1	1	1.000	0	1	0.000	3
hypothetical protein MGC2701	XM_414831	1	0	high	0	1	0.000	2
hypothetical protein MGC3520	XM_416285	2	1	2.000	0	2	0.000	4
aldehyde dehydrogenase 8A1	XM_419732	1	0	high	0	1	0.000	2
hypotheticalprotein MGC17943 LOC41806	NM_001006244	1	2	0.500	0	1	0.000	4
Rho GTPase-activating protei	XM_424446	1	1	1.000	1	0	high	3
immunity associated protein	XM_427236	4	1	4.000	2	2	1.000	6
NF-kappa-B-repressing factor NFKB	NM_001012887	1	0	high	0	1	0.000	2
growth arrest-specific protein	XM_415595	1	2	0.500	0	1	0.000	4
ABC transporter ABCG2 LOC42	XM_421638	1	1	1.000	0	1	0.000	3
matrix metallopeptidase7 matrilysin uterine	NM_001006278	2	1	2.000	0	2	0.000	4
AF15q14 protein isoform 2	XM_420937	1	0	high	0	1	0.000	2
ribosomal protein S11 RPS11	NM_001030833	3	4	0.750	2	1	2.000	8
KIAA2019 protein LOC423497	XM_421400	1	0	high	1	0	high	2
Peroxisome proliferator	XM_418125	1	0	high	1	0	high	2
hypothetical protein BC01514	XM_416964	1	0	high	0	1	0.000	2
4930438O03Rik protein LOC41	XM_415139	1	2	0.500	0	1	0.000	4
Voltage-dependent calcium	XM_415680	1	1	1.000	0	1	0.000	3
C6orf79 protein LOC420837	XM_418928	2	1	2.000	1	1	1.000	4
RIKEN cDNA 5830433M19 LOC42	XM_424940	1	0	high	0	1	0.000	2
mitotic spindle coiled-coil	XM_423216	1	0	high	0	1	0.000	2
retinoic acid receptor	XM_418471	1	0	high	0	1	0.000	2
KIAA0763 gene product LOC41	XM_414318	1	0	high	0	1	0.000	2
H2A histone family member V H2AFV	NM_001031374	1	2	0.500	0	1	0.000	4
solute carrier organic anion transporter family	NM_001030856	1	2	0.500	1	0	high	4
coiled-colLhelix-coiled-coil	XM_414369	1	0	high	0	1	0.000	2
acetoacetyl-CoA synthetase AACS	NM_001006184	4	2	2.000	1	3	0.333	7
follistatin-like 4 FSTL4	NM_204502	1	0	high	0	1	0.000	2
LOC416944 LOC416944	XM_429462	2	0	high	0	2	0.000	3

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
adaptor-related proteincomplex 3 mu 1 subunit A	NM_001031227	1	0	high	1	0	high	2
KIAA0893 protein LOC424344	XM_422187	1	1	1.000	1	0	high	3
dTDP-D-glucose 46-dehydratase	XM_416988	1	0	high	0	1	0.000	2
hypotheticalgene supported by BX930120	XM_430457	1	0	high	0	1	0.000	2
spermatogenesis associated 5	XM_413821	2	0	high	0	2	0.000	3
LOC419550 LOC419550	XM_429676	2	0	high	1	1	1.000	3
prekininogen LOC424957	XM_422766	4	3	1.333	1	3	0.333	8
FLJ45910 protein LOC426106	XM_423779	1	1	1.000	0	1	0.000	3
LOC407702 protein LOC420807	XM_418900	3	0	high	1	2	0.500	4
Chromobox protein homolog 7	XM_416253	1	0	high	1	0	high	2
bromodomain containing 9 LO	XM_418893	1	0	high	0	1	0.000	2
telomeric repeat binding factor NIMA-interacting	NM_204380	1	1	1.000	1	0	high	3
cystatin F; cystatin-like	XM_415013	2	1	2.000	0	2	0.000	4
NipSnap2 protein Glioblastoma amplified	NM_001030713	1	2	0.500	0	1	0.000	4
hypothetical protein LOC5106	XM_414741	1	1	1.000	0	1	0.000	3
S100 calcium binding protein A6 calcyclin S100A	NM_204148	1	0	high	1	0	high	2
cystatin C amyloid angiopathy and cerebral hemorr	NM_205500	5	0	high	4	1	4.000	6
hypothetical protein LOC1342	XM_424807	1	0	high	1	0	high	2
solute carrier family 38 member 2	NM_001030741	1	1	1.000	0	1	0.000	3
axonemal heavy chain dynein	XM_424606	2	0	high	1	1	1.000	3
cytochrome c oxidase subunit	XM_415270	2	3	0.667	0	2	0.000	6
RIKEN cDNA 2410043F08	XM_417617	2	1	2.000	1	1	1.000	4
myosin IF MYO1F	NM_205254	1	2	0.500	0	1	0.000	4
AQ LOC395744	NM_204914	5	1	5.000	3	2	1.500	7
hypotheticalgene supported by CR387606	XM_419472	1	0	high	0	1	0.000	2
tyrosine 3-monooxygenase/tryptophan 5-monooxygenas	NM_001006415	3	1	3.000	0	3	0.000	5
mKIAA1930 protein LOC415665	XM_414032	1	0	high	1	0	high	2
hypothetical protein LOC418	XM_417190	3	0	high	3	0	high	4
vacuolar protein sorting 45A yeast VPS45A	NM_001031593	1	2	0.500	0	1	0.000	4
MAWD binding protein Unknow	XM_421566	1	1	1.000	1	0	high	3
hypothetical protein FLJ1308	XM_428506	1	0	high	1	0	high	2
actin-related protein 10 homolog S. cerevisiae	NM_001006492	1	2	0.500	0	1	0.000	4
reserved; protein associatin	XM_422094	2	2	1.000	1	1	1.000	5
ectodysplasin A1 receptor associated death domain	NM_001012405	1	1	1.000	0	1	0.000	3
troponin C type 1 slow TNNC1	NM_205133	1	0	high	0	1	0.000	2
CNDP dipeptidase 2 metallopeptidase M20 family	NM_001006385	3	3	1.000	0	3	0.000	7
cysteine and glycine-rich protein 2 CSRP2	NM_205208	2	2	1.000	2	0	high	5
LOC426989 LOC426989	XM_430466	3	2	1.500	2	1	2.000	6
CG9147-PB LOC427125	XM_424718	2	2	1.000	2	0	high	5
hypotheticalgene supported by CR390948	XM_430202	1	0	high	1	0	high	2
hypotheticalgene supported by CR388998	XM_430449	3	1	3.000	0	3	0.000	5
ATP-binding cassette sub-family B MDR/TAP memb	NM_204894	1	2	0.500	1	0	high	4
RIM4 gamma LOC428142	XM_425700	2	2	1.000	0	2	0.000	5
CG10964-PA LOC425563	XM_423314	5	0	high	3	2	1.500	6
CG10964-PA LOC415663	XM_414030	2	1	2.000	1	1	1.000	4
hypothetical protein FLJ2263	XM_420923	1	1	1.000	1	0	high	3
hypothetical protein 3000008	XM_422572	3	3	1.000	2	1	2.000	7
LOC424715 LOC424715	XM_430177	2	0	high	0	2	0.000	3
KIAA1712; HBV PreS1	XM_420525	1	0	high	0	1	0.000	2
LOC425362 LOC425362 partial	XM_423133	3	2	1.500	1	2	0.500	6
interferon-related developmental regulator 1 IFRD	NM_001001468	2	1	2.000	1	1	1.000	4
ribosomal protein L21 LOC41	XM_417127	1	13	3.667	3	8	0.375	15
hypothetical protein FLJ2058	XM_421215	2	0	high	0	2	0.000	3
KIAA1462 protein LOC420476	XM_418578	4	1	4.000	1	3	0.333	6
chromosome 10 open reading frame	XM_424029	2	0	high	1	1	1.000	3
RIKEN cDNA 4930553M18 LOC41	XM_417196	1	1	1.000	0	1	0.000	3
septin 10 isoform 1 LOC4187	XM_416931	1	1	1.000	0	1	0.000	3
Neuron specific calcium-bind	XM_417815	1	1	1.000	1	0	high	3
LOC421159 LOC421159	XM_429816	3	0	high	1	2	0.500	4
reticulon 4 interacting protein	XM_419808	1	0	high	0	1	0.000	2
Downstream neighbor of Son	XM_416713	3	1	3.000	0	3	0.000	5
Testes development-related N	XM_424976	1	1	1.000	0	1	0.000	3

GenBank gene name	Accession Number	P_N	P_S	P_N/P_S	non-con	con	non-con/con	SNPs[1]
transmembrane gamma-carboxygenase	XM_419637	1	1	1.000	0	1	0.000	3
Nuclear protein SkiP Ski-in	XM_421294	1	2	0.500	1	0	high	4
Hypotheticalprotein KIAA0286 HA6800	NM_001031242	1	0	high	0	1	0.000	2
hypothetical protein LOC419	XM_417199	2	1	2.000	1	1	1.000	4
androgen-induced prostate proliferative	NM_001012827	1	0	high	1	0	high	2
hypothetical protein FLJ2050	XM_417759	3	3	1.000	1	2	0.500	7
hemoglobin alpha 2 HBA2	NM_001004376	26	5	5.200	8	1	8 0.444	32

Appendix C – Genes Associated with Susceptibility or Resistance to Disease

This list contains most chicken genes implicated in disease up to 2007. The disease abbreviations are listed in the table below:

Abbreviation	Disease	Abbreviation	Disease
SE	*Salmonella enterica* serovar Enteridis	SAT	spontaneous autoimmune thyroiditis
EC	*Escherichia coli*	HP	*Haemophilus paragallinarum*
ST	*Salmonella enterica* serovar Typhimurium	VSV	vesicular stomatitis virus
MDV	Marek's Disease Virus	DHBV	Duck Hepatitis B Virus
ALV	Avian Leukosis Virus	BA	*Brucella abortus*
RSV	Rous Sarcoma Virus	LM	*Listeria monocytogenes*
SA	*Staphylococcus aureus*	SRBC	Sheep red blood cells (an antigen)
(vv)IBDV	(very virulent) Infectious Bursal Disease Virus	EA	*Eimeria acervulina*
SG	*Salmonella gallinarum*	CP	*Clostridum perfringens*

Chicken genes implicated in disease:

Name(s)	Short name	GenBank	Other	Avian Diseases	Association Details	Publication References
		Accession Numbers				
ADL0146, ADL0355				EC	Novel region	Yunis et al '02
ADL0293, ADL0301				SE, EC	Novel region	
Alpha-enolase	ENO1		NM_205120	MDV	Novel gene	Niikura et al '04
Avian B-defensin-1	AVBD1	AF033335	NM_204993	Campylobacter, SE	Novel gene	Zhao et al '01; Lalmanach et al '06; Sadeyen et al '04
Avian B-defensin-2	AVBD2	AF033336	NM_204992	Campylobacter, SE	Novel gene	
Avian B-defensin-3	AVBD3	AY621318	NM_204650	Campylobacter, SE	Novel gene	Zhao et al '01; Hasenstein et al '06
Avian B-defensin-4	AvBD4	AY534892	NM_001001610	SE, ST	Antimicrobial action	Lynn et al '04, Milona et al '07
Avian B-defensin-7	AVBD7	AY534895	NM_001001194	SE, ST	Antimicrobial action	Lynn et al '04; Hasenstein et al '06, Milona et al '07
Avian B-defensin-8	AVBD8	AY534896	NM_001001609	SE, LM, ST	Aas causing changes in peptide charge	Higgs et al '07
Avian B-defensin-9	AvBD9	AY534897	NM_001001611	Campylobacter, ST, EC,CP, SE	Antimicrobial action	Lynn et al '04, Dijk et al '07, Milona et al '07
Name(s)	Short name	GenBank	Other	Avian Diseases	Association Details	Publication References
Avian B-defensin-11	AVBD11	AY621313, AY701473	NM_001001779			Lynn et al '04
Avian B-defensin-12	AvBD12	AY701474	NM_001001607	SE	SNPs	Lynn et al '05, Hasenstein et al '07
Avian B-defensin-13	AvBD13		NM_001001780	SE	SNPs	Hasenstein et al '07
Avian endogeneous viruses	EV-1	DQ118701		SAT	Novel gene	Vasicek et al '01
Avian	EV-3	CB016682		SAT	Novel gene	

endogeneous viruses						
Avian endogeneous viruses	EV-6			SAT	Novel gene	
B-F (region of MHC B complex)				MDV resistance	Novel genes	Briles et al '83
B13 (region of MHC)				MDV susceptability	Novel gene	Macklin et al '02
B2 (region of MHC)				MDV, ALV, SA	Novel gene	Longenecker et al '77; White et al '94; Cotter et al '92
B21 (region of MHC)				MDV resistance	Novel gene	Longenecker et al '77; Macklin et al '02
B8a (region of MHC)				ALV	Novel gene	Yoo et al '92
B9a (region of MHC)				ALV	Novel gene	
Blb (MHC class II beta gene)				MDV	Novel gene	Niikura et al '04
BQ (region of MHC)				SA	Novel gene	Cotter et al '02
BR3 (region of MHC)				RSV	Novel gene	White et al '94
BR4 (region of MHC)				RSV	Novel gene	
BR7 (region of MHC)				RSV	Novel gene	
Caspase-1	CASP-1	AF031351	NM_204924	SE	C/T at -368 bp of 5' flanking region	Liu & Lamont '03; Kramer et al '03; Ye et al '06
CCL-11				1918 influenza	Required immune response	Kobasa et al '07
Monocyte chemotactic protein-1	CCL-2			1918 influenza	Required immune response	
CCL-5	RANTES			1918 influenza	Required immune response	
CCLi4/MIP-1beta		AJ243034		TLR agonists	Differential response	Kogut et al '07
Name(s)	Short name	GenBank	Other	Avian Diseases	Association Details	Publication References
CD14				*Salmonella* LPS	Activates phospholipases in response	He et al '06
T-cell specific surface protein	CD28		NM_205311	SE	Novel gene	Malek et al '04
Chicken B-cell marker	ChB6	X92865		BA	Elevated antibody response	Zhou & Lamont '03
Chicken intestinal antimicrobial peptides	CIAMPs			IDBV	Enhanced immune response	Yurong et al '06
Complement component Clq-binding protein	C1QBP			MDV	Novel gene	Niikura et al '04
CXCL-11				1918 influenza	Required immune response	Kobasa et al '07
CXCLi1/K60		AF277660		TLR agonists	Differential response	Kogut et al '07
Four unnamed microsatellites				SE	Novel region	Kaiser & Lamont '02
Growth hormone	GH-1		NM_204359	MDV	Novel gene	Liu et al '01
GMCSF				SE	Differential response	Lalmanach & Lantier '99
Growth-related translationally	TPT1		NM_205398	MDV	Novel gene	Niikura et al '04

controlled tumor protein						
Guanylate cyclase	GC1	AF03942		Retinal degeneration	Null mutation	Semple-Rowland et al '88
Heterophils in general				SE resistance	Novel gene	Kogut et al '94

Inhibitor of apoptosis protein-1	IAP-1	AF221083;
		AF008592;
		AY494054
				SE	G/A = Ala at 157; C/T; elevated antibody response	Lamont et al '02; Zhou & Lamont '03; Liu & Lamont '03; Kramer et al '03; Ye et al '06

IFN-alpha		U07868		TLR agonists	Differential response	Kogut et al '06
IFNG		Y07922	NM_205149	SE, VSV, DHBV, BA	A/G at -318 bp of 5' flanking region; G/A; in promoter region	Kaiser et al '98; Zhou et al '01; Kramer et al '03; Okamura et al '04; Sadeyen et al '04; Long et al '05; Kogut et al '05; Ye et al '06
Ig L		M24403		SE	A/G at 60 bp upstream of octamer seq	Kramer et al '03
Name(s)	Short name	GenBank	Other	Avian Diseases	Association Details	Publication References
IL1B		AJ245728; Y15006	NM_204524	SE, Campylobacter	IR association	Kogut et al' 03; Smith et al '04; Okamura et al '04; Kogut et al '06
IL10				SE	Differential response	Lalmanach & Lantier '99, Ghebremicael et al '08
IL12				SE	Differential expression; induces IFN-g p4- subunit mainly	Sadeyen et al '04; DeJong et al '06
IL15R alpha		AI980376			Elevated antibody response	Zhou & Lamont '03
IL18		AJ416937	NM_204608	SE	Differential expression; induces IFN-g	Swaggerty et al '04; Kogut et al '06
IL2		AJ224516	NM_204153	SE, IBDV	A/G at -425 bp of 5' flanking region; A/C	Kramer et al '03; Kogut et al '03; Okamura et al '04; Li et al '04; Ye et al '06, Tarpey et al '07
IL4				SE	Differential expression - susceptability	Smith et al '04; Lalmanach & Lantier '99
IL6		AJ309540; AJ250838		SE, Campylobacter, IBDV	Differential expression - susceptability	Lalmanach & Lantier '99; Kogut et al '03; Smith et al '04; Okamura et al '04; Swaggerty et al '04; Sun et al '05; Kogut et al '06; Kobasa et al '07
IL8 (previously known as CXCLi2 and CAF)		AJ009800		SE, Campylobacter, TLR agonists, 1918 influenza	Differential expression - susceptability	Kogut et al '02; Swaggerty et al '03; Enocksson et al '04, Smith et al '04; Lalmanach et al '06; Kogut et al '06; Kobasa et al '07

Inducible Nitric oxide	INOS	D85422;
		U34045;
		AF537190;
		U46504
				SE, Campylobacter	C/T in intron; T/C	Enocksson et al '04; Smith et al '04; Kramer et al '03; Eisenstein '01; Lalmanach et al '06; Ye et al '06

Invariant (Ii) chain	CD74		NM_001001613	MDV	Novel gene	Niikura et al '04
K60 (a cytokine)		AF266770		Campylobacter	Novel gene	Smith et al '04
LEAP-2				ST (some strains)	Novel allele	Townes et al '04
L-meq				MDV	Novel gene	Chang et al '02
LEI0104			NW_060306	EC	Novel region	Yunis et al '02
LEI0135			NW_060670	SE, EC	Novel region	
Name(s)	Short name	GenBank	Other	Avian Diseases	Association Details	Publication References
Lymphocyte antigen 6 complex locus E, alias SCA2 or TSA1	LY6E			MDV	Novel gene	Liu et al '03
MCW0024				SE, EC	Novel region	Yunis et al '02
MCW0051			NW_060298	SE	Novel region	Yunis et al '02
MCW0083				SE, EC	Novel region	
MCW0083 II (adjacent to BMP2, bone morphogenetic protein 2)				SE, EC	Novel region	
MCW0183 III				EC	Novel region	
MCW0214				EC	Novel region	
MD-2 (accessory protein of TLR-4)				SE	G/A	Malek et al '04; Ye et al '06
Meq				MDV resistance	Novel gene	Chang et al '02
MHC ("K" classIV)				EA	Novel gene	Uni et al '95
MHC B15				SE resistance	Novel allele	Cotter et al '98
MHC class 1 alpha transcript 1.5 + 1.9				MDV resistance & susceptability	Novel transcript	Dalgaard et al '03
MHC class 1 B-FIV-B12 alpha-chain		M31012		SE	Lys->Met at 148	Lamont et al '02;
MHC class 1 alpha (2) domain		AF459826 - 30		SE	A/T = Lys->Met at 148	Liu et al '02
MHC generally				*Salmonella*, fowl cholera, coccidosis, RSV, MDV, helminthe parasites, ALV, fowl cholera	etc	Fulton et al '06; Kaufman et al '07; Schou et al '06;
Mim-1 (P33 = protein product)		M29449		SE	Differential expression	Bischoff et al '01; Crippen et al '03; Lalmanach et al '06
Name(s)	Short name	GenBank	Other	Avian Diseases	Association Details	Publication References
Myxovirus resistance	Mx	AB088536	NM_204609	Differential antiviral capacity, resistance & susceptability	Asn(r) to Ser(s) at 631; 2032 tcn->aa(t/c) - G to A	Ko et al '02; Li et al '06; Seyama et al '07; Xu et al '07; Ko et al '04, Balkisson et al '07
Natural Killer cell receptor (in MHC)	Nkr			SE resistance	Novel gene	Kautman et al '99
Natural resistance-associated macrophage protein 1 (NRAMP1)	SCL11A1	U40598; AY072001	NM_204964	SE	C/T; Ser 379; G/A at 696 Arg to Gln 223 in TM5-6 region	Hu et al '97; Govoni et al '98; Lamont et al '02; Girard-Santosuosso et al '02; Liu et al '03; Kramer et al '03; Beaumont et al '03; Sadeyen et al

						'04, Wigley 2004
Ovotransferrin				MDV; viruses generally		Giansanti et al '07
Prosaposin	PSAP	AB003471	NM_204811	SE	G/A; Gly 271	Lamont et al '02; Liu & Lamont '03; Kramer et al '03
Retinoblastoma-binding protein4	RBBP4		NM_204852	MDV	Novel gene	Niikura et al '04
Rfp-Y (MHC-like region)				MDV resistance	Novel gene	Lakshmanan et al '97
Sal 1				SE resistance	Novel gene	Wigley et al '02; Mariani et al '01, Wigley 2004
SC 1		S63276		Nephroblastomas	Novel gene	Tsukamoto et al '05
TcR or CD28				Schleroderma		Zekarias et al '01
Transforming growth factor Beta 2	TGF-beta 2	X58071	NM_001031045	SE	C/T at -1667 bp of 5' flanking region	Kramer et al '03
Transforming growth factor Beta 3	TGF-beta 3	X60091	NM_205454	SE	C/A at 2833; C/T at -171 bp of exon 5; C/A; T/C	Malek & Lamont '03; Kramer et al '03; Ye et al '06
Transforming growth factor Beta 4	TGF-beta 4	AF459837, M31160		SE	A/C at Glu-> Asp at 210	Kramer et al '03; Swaggerty et al '04 (-ve association)
Toll-like receptor 4	TLR4	AY064697	NM_001030693	SE	C/A; activates phospholipases (along with CD14) in response to Sal. Lps	Beaumont et al '03; Leveque et al '03; Malek et al '04; Lalmanach et al '06; Ye et al '06; He et al '06
Tenascin C (Lps/Tlr4)	TNC		NM_205456	SE	Novel gene	Hu et al '97
Name(s)	Short name	GenBank	Other	Avian Diseases	Association Details	Publication References
Tumour necrosis factor alpha	TNF-A			SE	Differential expression	Lalmanach & Lantier '99; Lynn et al '03
TNF related apoptosis inducing ligand	TRAIL	AB114678; AF537189		SE	G/A at 82 bp; A/G	Malek & Lamont '03; Ye et al '06
Tva locus				ALV, ASV (both type -A)	Allele for receptor for virus	Bates et al '98
Tvb locus	Car1			ALV (type -B & -D)	Allele for receptor for virus	Smith et al '98
ZOV3		AF221566; D16151		SE	A/G; Val->Leu at 216; elevated antibody response	Kramer et al '03; Zhou & Lamont '03