DETECTION OF SELECTION IN MAMMALIAN GENOMES AND POPULATIONS

A dissertation submitted to the University of Dublin
for the Degree of Doctor of Philosophy
2010

Lilian Pek Lian Lau

School of Genetics and Microbiology
Trinity College Dublin

Declaration

I hereby certify that this thesis has not been previously submitted for examination to this or any other university. The work described herein is entirely my own work (except where indicated).

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.

Summary

Genome-wide scans for evidence of positive selection in mammalian genomes have recently become possible with the availability of whole genome sequences. They offer a chance to identify those genes that were of most importance during evolution and which have undergone adaptive change. It has also been hypothesised that identification of these signatures may reveal the genes that engender the key biological differences between a species of interest and other related species.

The primary objective of this thesis was the detection of signatures of selection in the bovine genome and in cattle populations. Since their domestication approximately ten thousand years ago, cattle have been selected for breeding based on a range of desirable traits, including milk yield, beef quality, draught ability, temperament and disease resistance/tolerance.

The principal method of detecting genes positively selected between different species is to carry out a comparative analysis of the rate of nonsynonymous (protein-changing) to synonymous (silent) changes (d_N/d_S). This method may be conservative, however, in that it requires the d_N/d_S ratio to be elevated over an entire gene while it is more usual for positive selection to act only on particular codons or domains. To overcome this limitation, we have developed an approach to systematically examine gene regions which encode extracellular domains for evidence of positive selection in the human, chimpanzee and bovine genomes. I postulated that selection is likely to be more prevalent in such domains due to direct interactions between extracellular regions of proteins and the external environment. Several, interesting candidate genes were identified in this thesis using this approach, which were not identified in several other analyses of signatures of adaptive evolution in these genomes. I also present work from our involvement in the international bovine genome project to identify genes which have putatively been subject to positive selection.

This thesis also presents an approach to utilise Expressed Sequence Tags (ESTs) to identify and survey functionally relevant polymorphisms in bovine populations. The development of a whole genome SNP genotyping platform for cattle, however, has allowed us to investigate signatures of domestication and artificial and natural selection not only in the bovine genome in comparison to other species but also within bovine populations. Several approaches were used in this thesis to detect selection in cattle populations, including (a) the F_{ST} statistic, a measure of the proportion of genetic variance in a subpopulation relative to the total genetic variance, (b) Locus Specific Branch Length (LSBL), an extension of the F_{ST} statistic used to investigate intercontinental relationships between the bovine populations studied, (c) Fay & Wu's H statistic, which utilise derived and ancestral allele information, and (d) composite empirical calculations, which enabled identification of candidate genes with greater precision. This work also formed our contribution to the international bovine HapMap project.

Three geographically distinct cattle populations were studied – European *Bos taurus*, African *Bos taurus* and (Indian) *Bos indicus*. Outgroup species including yak, gaur, bison, water buffalo and anoa were used to infer ancestral and derived alleles. Several potential regions under selection were identified; including genes associated with disease and immunity, milk processing, development, muscling and stature. We argue that the results from these disparate approaches indicate the power of whole-genome analysis to identify traits of economic and historical importance.

Acknowledgements

"It would be so nice if something made sense for a change."

"Curiouser and curiouser."

-- *Alice in Wonderland*

I have no word of thanks that can sufficiently express my gratitude to Dave and Dan. Dave has been with me every step of the way, not only as a mentor but also a friend. He has been more than patient with me - particularly when I veered off course and needed help to get back on track, taught me more about science, research, networking and life than anyone I can think of right now, has been engaged and encouraging with my work, and most of all, having faith in me when I have little to motivate me on. Dan has been an invaluable advisor and understanding supervisor, with a wealth of knowledge and information to impart like little nuggets of gold. Truly, without both Dave and Dan, this thesis would not have been possible.

To the members of the lab, past and present, including Valeria, Emma, Brian, Ruth, Ceiridwen, Caitríona, Tim, Frauke, Sarah, Yonas, Matthew, Russell and Kevin, thank you for every little help, tips and tricks, discussions, goodies and treats, lab lunches/ drinks/ birthdays/ celebrations, and more importantly, the support since day one.

I would like to also thank Prof. Fiona Brinkman for hosting my short research stay in Simon Fraser University, Vancouver. I am also appreciative of the warm welcome that I've received from the members of her lab - Ray, Matthew L, Matthew W, Geoff, Nancy, William, Nicolas, Amber, Morgan, Mark, Tim and Chi Kin.

I am also appreciative for the opportunity to be involved in both the international bovine genome and bovine HapMap projects, and for the datasets provided by both consortiums.

This journey would have been a lonely one to embark on had I not have my friends who are constant source of company, encouragement and advice. With friends living locally and abroad, there's always someone to talk to (instant messaging), to ask help from (work-related or not, usually not), or to just say hello to, regardless of the time. Sila, Nora, Anne-Laure, Lisa, Åsa, Naisha, Else, Miriam, Efi, Kevin, Gavin, Marco, Kim, Hui Mei, Rachel, Siang Lee, Jin Fen, Jee Ming, Daniel, Li Lian, Vincent, Michael G, Michael E, Julien, I could go on. Thank you everyone and I'll be back following your status updates before you know it.

I'm quite convinced that Jorge Cham has a hidden camera that follows me around. The PHD comics keep me entertained and sane, while its Grad Forum (a.k.a. Phorum) has been a tremendous network of support and knowledge sharing across the globe. I have also made some wonderful friends and foremost among them are Chloé and Anne who have welcomed me into their world, complete with their families and friends. Paris is nought but a geographical separation, and we'll have to embark on more foodie adventures soon - chocolates and macarons awaiting!

To my family, I thank them for their unconditional love and support, even when many of them may not fully understand what I've been and am doing ("too Science-y"). The family has generations of chefs, entrepreneurs, accountants and healers but not scientists, making this an uncharted territory. Food is what this family does best, and it shows. I get regular stocking up of home-cooked meals, if not treated with the best that restaurants in Dublin have to offer, from my younger brother Andrew and my sister-in-law Sook Lee, and my aunts and uncles - Alice, TK, Janice, Dermot, Judy, Laurence. I was probably the best fed grad student in the city. My youngest brother Beejay, home in Malaysia, keeps me updated with regular news and photographs of my mum and the rest of my family at home, including my wonderful grandparents who believe in the values of education and hard work. To my cousin Serena, who is as good as a sister (and the only other scientist of the family for now), thanks for listening to all the Science-speak.

This thesis is dedicated in memory of:

Dermot P Burke (1919-2009) – uncle, friend and confidant; for always believing in me and showing me the beauty of arts and literature.

Chuan Hock Lau (1930-2010) – my beloved grandfather who was always there to love and to take care of everyone unconditionally; who taught me to be thankful for every little blessings and to be generous to others, be it of material, time and/or affection; who showed me just how brave he was in facing difficult times and even then, put others first ahead of himself.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AHMM	Extension of TM-HMM
BEB	Bayes Empirical Bayes
BGP	Bovine genome project
CDS	Coding sequence
d_N/d_S	Rate of nonsynonymous substitution to synonymous substitution
dbEST	Database of EST
EC	Extracellular (domain)
EHH	Extended haplotype heterozygosity
EST	Expressed sequence tag
FDR	False discovery rate
F_{ST}	Proportion of genetic variance in a subpopulation relative to the total genetic variance
GO	Gene ontology
HCM	Human, chimpanzee and mouse trio
HCR	Human, chimpanzee and rhesus macaque trio
HMRDC	Human, mouse, rat, dog and cow quintet
HMM	Hidden Markov Model
iHH	Integrated EHH
iHS	Integrated haplotype score
lnL	Log likelihood value
LRH	Long-range haplotype
LRT	Likelihood ratio test
LSBL	Locus specific branch length
MEGA	Molecular Evolutionary Genetic Analysis
mtDNA	Mitochondrial DNA
MWU	Mann-Whitney U test
PAML	Phylogenetic Analysis by Maximum Likelihood
PANTHER	Protein Analysis Through Evolutionary Relationship

PCR	Polymerase chain reaction
QTL	Quantitative trait loci
NEB	Naïve Empirical Bayes
nsSNP	Non-synonymous SNP
RefSeq	Reference sequence
SIFT	Sorting intolerant from tolerant
SNP	Single nucleotide polymorphism
synSNP	Synonymous SNP
TM-HMM	Protein topology prediction method based on a hidden Markov model
XP-EHH	Cross population extended haplotype heterozygosity

BGSAC	Bovine Genome Sequencing and Analysis Consortium
CSAC	Chimpanzee Sequencing and Analysis Consortium
RMGSAC	Rhesus Macaque Genome Sequencing and Analysis Consortium

1. General Introduction

1.1 Natural Selection

The phenomena of selection and mutation are central subjects in the understanding of molecular evolution. The neutral theory formulated in the 1960s claims that mutations that give rise to most variation do not affect the fitness of an organism. Instead, these variations are attributed to random fixation (Kimura 1968). Therefore, the observed evolution of neutral sites within sequences may be used for convenient inference of mutation rates. According to this theory, each generation would have $2N\mu$ new neutral mutations where μ is the neutral mutation rate, and N is the diploid population size. The probability of fixation of a new mutation is $1/(2N)$ in a given population. As a result, its rate of fixation is $(2N\mu)/(2N) = \mu$.

Subsequent studies, however, raised questions that cannot be answered by the neutral theory. For one, this theory predicts that species with small populations should show lower levels of polymorphism than those with large populations but the observed difference is modest. Moreover, it also predicts that species with shorter generation times would have faster evolving proteins but this is not always the case (Ohta and Gillespie 1996). This led to the development of the nearly neutral theory which considers mutation that does have a small effect on fitness (Ohta 1995). Under this model, a mutation is "effectively neutral" if its selective disadvantage is small in comparison with the population size (Kreitman 1996), such as that for mammals, or when the species are isolated on islands (Woolfit and Bromham 2005).

1.1.1 Positive Selection

Selectively advantageous mutations are those which confer beneficial traits in species survival and success in reproduction. These mutations spread rapidly and are maintained in a population by positive selection or adaptive evolution. A gene may be inferred to be subject to positive selection when its rate of nonsynonymous substitutions per nonsynonymous site (d_N) to synonymous substitutions per synonymous site (d_S), indicated by d_N/d_S or ω, is significantly greater than 1.

Nonsynonymous substitutions are amino acid changing mutations while synonymous substitutions are silent at the level of protein sequence, although there is recent emerging evidence for non-neutral evolution at synonymous sites (Chamary, Parmley, and Hurst 2006). An elevated d_N/d_S value, which is indicative of an accelerated rate of protein evolution, may be due to adaptive evolution or the relaxation of purifying selective constraints. While $d_N/d_S > 1$ indicates positive selection or adaptive evolution, $d_N/d_S < 1$ implies the action of negative purifying selection and $d_N/d_S = 1$ indicates neutral evolution (Yang 2002).

1.1.2 Interspecies Detection of Positive Selection

There are numerous methods to detect interspecies events of positive selection, and some of the most commonly used methods are discussed below.

1.1.2.1 Early Studies

Early studies to detect positive selection were carried out using pairwise sequence analysis which tested if $d_N - d_S$ was significantly greater than 0 (e.g. (Endo, Ikeo, and Gojobori 1996). This approach, however, lacked power to detect positive selection. It averages the substitution rate over all amino acid sites over time, which is unrealistic given that most sites are expected to be highly conserved with selection occurring only on a few sites at certain time points. To avoid averaging rates over long periods of time, Messier and Stewart (Messier and Stewart 1997) calculated d_N and d_S for each branch in a phylogeny. They were able to detect positive selection along particular branches by investigating whether $d_N > d_S$, and normally approximated it to the statistic $d_N - d_S$.

While the method was an improvement, there were concerns regarding the reliability of the normal approximation. Zhang *et al.* suggested the application of Fisher's exact test to the counts of differences (Zhang, Kumar, and Nei 1997). Nonetheless, the estimates of substitution rates along every branch were relatively simplistic without taking account of sequence evolutionary features, such as transition/transversion rates and codon usage bias. Codon usage bias, for example, alters the frequency of occurrence of synonymous codons.

1.1.2.2 Maximum Likelihood Method

One method that is now frequently used to calculate the d_N/d_S ratio is to fit models of evolution by maximum likelihood to coding sequence alignments from different species. Several models which utilise maximum likelihood have been developed, and are implemented in packages such as Phylogenetic Analysis by Maximum Likelihood (PAML). Branch models allow the ratio of d_N/d_S to vary among branches to detect positive selection acting on particular lineages. They are, however, considered to be conservative, requiring $d_N/d_S > 1$ over the entire gene as it assumes all amino acid sites to be under the same selective pressure (Yang 2002).

Site models allow d_N/d_S to vary among codon sites (Nielsen and Yang 1998). The site detection is localised using an empirical Bayesian method, with the parameters used to calculate posterior probabilities estimated by maximum likelihood. Earlier implementations of empirical Bayes which did not account for uncertainty in the estimates of the parameters is known as naïve empirical Bayes (NEB) while recent approaches that do address the uncertainty to some extent is known as Bayes Empirical Bayes (BEB). Comparisons between the NEB and BEB to identify selected sites suggest that accounting for the uncertainty in parameter estimations is important for accuracy in detecting selected sites (Huelsenbeck and Dyer 2004; Yang, Wong, and Nielsen 2005).

Branch-site models allow d_N/d_S to vary among both sites and lineages to enable the detection of positive selection that affects only a few sites along a few lineages (Yang and Nielsen 2002). However, following simulation studies, Zhang found that the branch-site models are unable to distinguish between relaxation of selective constraints and positive selection (Zhang 2004). A modified branch-site model was subsequently developed to address this shortcoming (Zhang, Nielsen, and Yang 2005).

The log likelihood (lnL) values estimated by the maximum likelihood models are used to calculate a likelihood ratio test (LRT) of the hypothesis tested. The LRT is given by twice the difference of the lnL values of the two models tested, denoted by $2\Delta\ell$. The LRT value is subsequently compared to a χ^2 distribution to determine which model is statistically favoured.

1.1.2.3 Comparing Total Substitution Sites

There are several other approaches to infer positive selection, particularly at single amino acid sites. Fitch et al. (Fitch et al. 1997) used multiple sequence alignments of coding sequences to construct the most parsimonious phylogenetic tree, i.e. one that has the minimum number of evolutionary changes. For every codon site throughout the tree, the total number of nonsynonymous substitutions was compared to the total number of synonymous substitutions to identify (hypervariable) sites which have a greater proportion of nonsynonymous changes than the average over the sequence.

1.1.2.4 Parsimony Methods

Suzuki and Gojobori (Suzuki and Gojobori 1999) also applied a parsimony method to detect positive selection at single amino acid sites. Using multiple sequence alignments, the number of synonymous substitutions was used to reconstruct a phylogenetic tree by neighbour-joining (Saitou and Nei 1987). For each codon site, the ancestral codon at each node of the phylogenetic tree was inferred. The average numbers of synonymous and nonsynonymous sites were then estimated at each site, and the total numbers of these changes were also counted. The proportion of nonsynonymous changes was then calculated.

Under the null hypothesis, $d_N = d_S$ and positive selection is inferred when the null hypothesis is rejected i.e. $d_N > d_S$. Simulation studies have shown maximum parsimony methods to be less reliable than maximum likelihood, particularly when sequences are divergent (Zhang and Nei 1997). On the other hand, when the sequences compared are closely related to each other, then both methods produce relatively reliable results. It is also of note that maximum parsimony requires large numbers of sequences to derive significant results.

1.2 Comparative Genomics and the Study of Positive Selection

In recent years, whole genome sequencing efforts have brought about the availability of sequences from different species and with it, the ability to systematically conduct comparative genomics studies of multiple species of interest. The underlying approach to a comparative genomics study is the alignment of two or more genomic sequences, the identification of orthologous (or paralogous) genes, and subsequent analyses to determine the extent of sequence conservation and divergence of the genes. The identification of signatures of adaptive evolution (i.e. sequences that evolve beyond neutral expectation) is of particular interest to our work. In this thesis, I present the comparative analysis of several mammalian genomes including those of human (Lander et al. 2001; Venter et al. 2001), chimpanzee (CSAC 2005), macaque (Gibbs et al. 2007), mouse (Waterston et al. 2002), rat (Gibbs et al. 2004), dog (Lindblad-Toh et al. 2005) and cow (Elsik et al. 2009).

1.2.1 The Study of Positive Selection in Pre-Genomics Era

Prior to the availability of genome sequences, studies of positive selection had been focused on single candidate-gene analyses. Such studies require a hypothesis about which genes may be subject to positive selection. A detailed understanding of the gene's functions was an important criterion in proposing likely candidates for investigation. Genes involved in processes such as the immune response, reproductive system, or olfaction, generally provided the mostly likely candidates for studies of positive selection. Such studies are, however, biased away from the identification of poorly characterised genes or genes unsuspected as being candidates of selection.

Despite the limitation of this approach, it has yielded some notable success stories. A mutation of the regulatory region near the human lactase (LCT) gene (Scrimshaw and Murray 1988), for example, has been shown to be highly selected in north central Europe. This mutation is postulated to confer lactase persistence in adulthood as a result of selective pressure following cattle domestication and the introduction of dairy products into the diet (Beja-Pereira et al. 2003).

Another gene that has been identified to be under positive selection in humans is the Duffy antigen (FY) gene which is linked to resistance to *Plasmodium* (malaria). The gene encodes a membrane protein that is exploited as an entry point for malaria parasites to enter red blood cells. A mutation in the promoter region of this gene disrupts gene expression and provides a protective mechanism against parasite entry. The allele frequency of this mutation is 100% in some regions of western Africa, where malaria has been endemic (Tournamille et al. 1995).

1.2.2 Comparative Genomics Approach

Comparative genomics allows a genome-wide investigation of signatures of selection without prior postulation of the type of gene to focus the analysis on. While this approach is most effective with complete or nearly complete genomes, it is also applicable to any large-scale gene datasets generated by expressed sequence tag (EST) or shot-gun sequencing.

Many of the studies to date have focused on positive selection on the human lineage (and by extension, the primate lineages) against a backdrop of other lineages. Clark *et al.*, for example, searched for signatures of positive selection using orthologous trios of human, chimpanzee and mouse sequences (Clark et al. 2003) with a focus on selection on human and chimpanzee lineages. With similar interest, Nielsen *et al.* carried out a pairwise analysis of human and chimpanzee genes (Nielsen et al. 2005) but this method is limited, as it cannot distinguish on which lineage experienced positive selection. The Chimpanzee Sequencing and Analysis Consortium similarly analysed human-chimpanzee orthologues, and they also carried out further analysis with additional mouse and rat sequences to differentiate between the hominid and the murid lineages (CSAC 2005).

In analysing the canine genome, Lindblad-Toh *et al.* focused on comparative studies of human, mouse and dog orthologues (Lindblad-Toh et al. 2005). Taking advantage of the new availability of the canine dataset, Arbiza *et al.* used human, chimpanzee, mouse, rat and dog to elucidate cases of positive selection in the primate lineages. The murids and the dog were used as weighted outgroups (Arbiza, Dopazo, and Dopazo 2006). This study employed the newly improved maximum likelihood branch-site model to differentiate relaxation of selective constraint from positive selection.

Bakewell *et al.* investigated the genomes of human, chimpanzee and macaque (Bakewell, Shi, and Zhang 2007) following the availability of the macaque genome. Macaque is a closely-related species and serves as a more suitable outgroup than previously used species including the rodents and dog. Kosiol *et al.* extended their study further to include human, chimpanzee, macaque, mouse, rat and dog sequences (Kosiol et al. 2008).

These comparative analyses enabled the genome-wide identification of genes that have putatively been subject to positive selection. Several categories of genes were enriched for signatures of adaptive evolution including those involved in immunity and reproduction; the former is likely driven by an arms race against invading pathogens and the latter by competition for fertilisation success. A number of other functional categories were also found to be enriched in genes under positive selection including those involved in sensory perception, apoptosis, signal transduction and cell adhesion.

1.3 Selection in Populations

Selection within populations occurs in a much shorter evolutionary history compared to selection between species. Population-level selection results in notable changes in genetic diversity and allele frequencies in regions of the genome subject to selection. Such "signatures" of selection may be detected using a variety of population genetics tests (discussed below). Given that most genetic variation evolves mainly under a neutrality model, regions with signatures of selection can most reliably be identified through comparisons to the genome-wide distribution of genetic variation.

1.3.1 Types of Selection in Populations

Three types of selection occur in populations – positive selection, negative selection, and balancing selection. Selection history at a given locus may also include an amalgam of these, integrated over time. Positive selection is also known as directional selection or Darwinian selection, and refers to the selective process where advantageous alleles are driven to high frequency, and in cases of strong selection – to fixation, in populations. The allele frequency spectrum for positive selection skews towards low frequency alleles, with an excess of high-frequency derived alleles (Bamshad and Wooding 2003) (Figure 1.1). Negative or purifying selection is the process that removes deleterious mutations. This is the most common form of natural selection.

Balancing selection is selection that favours variability/diversity within populations (Lewontin and Hubby 1966). Also known as disruptive selection, it maintains multiple alleles above neutral expectations, although the spread of the alleles never reaches fixation. Balancing selection shifts the allele frequency spectrum towards an excess of intermediate frequency alleles (Figure 1.1).

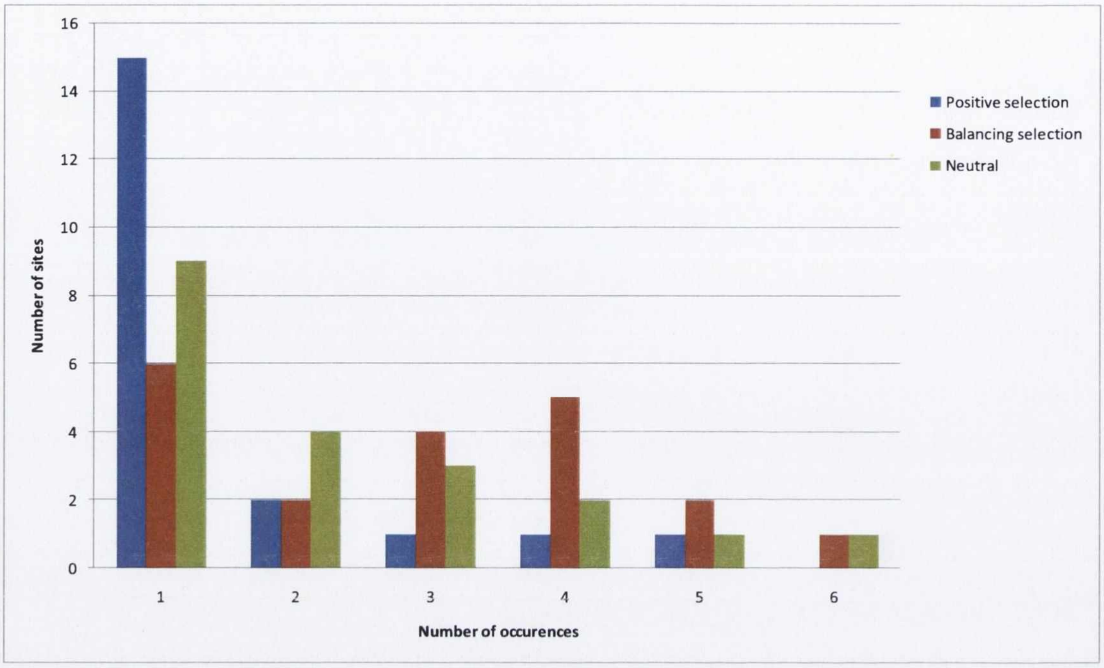


Figure 1.1 | Bar chart to illustrate the site frequency spectrum, for example gene with 20 segregating sites, under positive selection, balancing selection and neutral model. Positive selection skews the allele frequency spectrum towards lower level of sequency diversity and an excess of low-frequency variants. Balancing selection results in higher level of sequency diversity with a distribution reflecting an excess of intermediate-frequency variants.

1.3.2 Methods to Detect Positive Selection in Populations

The detection of selection within a population requires different statistical approaches depending on the nature of the selection and the window of evolutionary time that selection has taken place. Single nucleotide polymorphisms (SNPs) are the most common type of genetic variant from which information can be mined to elucidate the events of selection.

There are broadly four categories of signatures of positive selection within species: (i) the reduction in genetic diversity with an excess of rare alleles, (ii) a high-frequency of derived alleles, (iii) differences between populations, and (iv) long haplotypes.

1.3.2.1 Reduced Diversity with Excess of Rare Alleles

One signature of positive selection is a reduction in genetic diversity due to a hitchhiking effect. When an allele increases in population frequency, the linked variants in neighbouring locations also rise in frequency, leading to a "selective sweep". When a complete selective sweep takes place, it brings the selected allele to fixation and with it, the associated alleles too. As a result, the genetic diversity in close vicinity to the selected allele is eliminated and is reduced in an extended region surrounding the allele. With time, new mutations introduce diversity into the region but the process is slow and these variants are initially at low frequency.

The most commonly used tests to detect regions of low diversity with an excess of rare alleles include Tajima's *D*, and Fu and Li's *D* and *F*. Tajima's *D* tests involve comparison of the heterozygosity, π (i.e. average number of pairwise differences between two samples), and the number of segregating sites, *S*. Negative values of Tajima's *D* indicate an excess of rare alleles (Tajima 1989). Fu and Li's tests incorporate information from outgroup species and measure the number of variants observed only once in the sample versus either the total number of variant sites (Fu and Li's *D*) or π (Fu and Li's *F*). An excess in relative singleton variants is indicative of positive selection (Fu and Li 1993).

Identifying reduced diversity can be particularly useful because of its persistence in the genome compared to other population genetic signatures. For a new mutation to drift to higher frequency under neutrality, in human populations, approximately 1 million years is required (Sabeti et al. 2006). Therefore a statistically significant signal can persist for hundreds of thousands of years. However, it is challenging to distinguish this signature from the effects of population history - following a population expansion, for example, there is a similar increase in the fraction of rare alleles.

1.3.2.2 High-Frequency Derived Allele

Another signature of positive selection is an excess of derived alleles. Non-ancestral alleles arise from new mutations and are typically lower in frequency than ancestral alleles. However, following a selective sweep, derived alleles that are linked to the selected allele can rise to high frequency and are subsequently maintained in the population. A signature of high derived allele frequency is therefore created from an event of positive selection.

A derived allele signature is different from the excess of rare alleles signature. While both variants hitchhike to high frequencies, demographic confounders that affect the signatures are different. Rare alleles are mostly affected by population expansion but not the derived alleles, whereas population subdivision is more of a problem in detecting derived alleles (Przeworski 2002).

The most commonly used test for derived allele excess is Fay and Wu's *H* (Fay and Wu 2000). To carry out this test, it is imperative that the ancestral alleles are known and this can be inferred with the inclusion of an outgroup sequence (of a closely related species). It is also assumed that the mutation occurred only once at a particular locus and that the mutation took place after the species divergence.

1.3.2.3 Differences between Populations

When different populations of a species are geographically separated, each may be subject to diverse selective pressures, commonly dictated by environment and perhaps culture. Therefore selection that acts on one population may not act on another. A relatively large difference in allele frequency at a particular gene may signal a locus which has undergone positive selection.

A commonly used statistical test to measure divergence between populations is F_{ST}, a measure of the proportion of genetic variance in a subpopulation relative to the total genetic variance (Akey et al. 2002). Positive selection that takes place in one population but not the other drives the selected loci to an increase in F_{ST}. The range of the value of F_{ST} is 0 to 1, where value of 1 denotes completely divergent allele frequencies and value of 0 indicates complete lack of divergence between the populations tested.

One issue in detecting positive selection using F_{ST} statistics is in distinguishing between selection and the effect of demographic history. Events, such as population bottlenecks, may reduce diversity in one population but not the other, creating significant differences in pairwise comparisons.

1.3.2.4 Long Haplotypes

When a selected allele under positive selection rises to fixation, it may be at a rate that prohibits substantial recombination, thus forming long-range homogenous haplotypes. This selective sweep produces alleles that are both at high frequency and have long-range associations with other alleles.

This signature is commonly detected using the Long-Range Haplotype (LRH) test (Sabeti et al. 2002), the integrated Haplotype Score (iHS) (Voight et al. 2006), or the Cross Population Extended Haplotype Heterozygosity (XP-EHH) test (Sabeti et al. 2007). The LRH test is conducted based on the selection of a "core" haplotype and the decay of linkage disequilibrium (LD) on flanking markers is assessed by the calculation of the extended heterozygosity haplotype (EHH). A core haplotype refers to the haplotype at a locus of interest and it may consist of a single SNP.

Using LRH, evidence of positive selection is tested by detecting core haplotypes with elevated EHH relative to other core haplotypes. The iHS test computes the integrated EHH (iHH) and large positive and negative values of iHS indicate long haplotypes for the ancestral and the derived allele, respectively. The XP-EHH approach allows the elucidation of selective sweeps whose selected allele has reached or is approaching fixation (allele frequency ~80-100%) in one population, but remains polymorphic in the test population as a whole.

Long haplotypes are particularly useful in detecting partial selective sweeps, and the tests for this signature are relatively robust to datasets containing ascertainment bias. Moreover, the tests allow narrowing down of candidate regions, even down to a single gene. The limitation in long-range haplotypes as a signature of selection is that such haplotypes persist only for a limited time period (typically less than 30,000 years) as recombination breaks the region down, leaving fragments too short for detection.

1.4 Genome-Wide Detection of Selection in Populations

Prior to the genomics era, investigations of which genes were subject to selection were carried out through candidate gene analyses. While these studies have identified a number of genes under selection, such as the signature of positive selection on the LCT gene, which likely led to the persistence in lactose tolerance in human populations (Bersaglieri et al. 2004), this strategy is limited in that it requires an *a priori* hypothesis of which genes may have been selected. Such studies are biased in that only genes with a likely assumption of being subject to selective pressure are investigated. Moreover, given the confounding affects of population demographic history and selective pressure, it is difficult to interpret patterns of genetic variation for any single locus as indicative of positive selection.

1.4.1 Genome-Wide Studies

Genome-wide studies of positive selection are now possible, thanks to the availability of genome sequences particularly human (Lander et al. 2001; Venter et al. 2001), chimpanzee (CSAC 2005), mouse (Waterston et al. 2002), dog (Lindblad-Toh et al.

2005) and cow (Elsik et al. 2009)[1], and the availability of genetic variation datasets such as Haplotype Map (HapMap) SNP datasets. Large genome-wide SNP datasets are now available for human (Gibbs et al. 2005; Frazer et al. 2007a), rat (Saar et al. 2008), chicken (Wong et al. 2004) and cattle (Gibbs et al. 2009)[2]. The term "population genomics" has been coined to refer to the inference of population genetic and evolutionary parameters based on genome-wide datasets.

The current data, however, are still limited. Improved SNP coverage and improved genome assemblies and annotation are required for most of species (e.g. the current bovine SNP chip contains approximately 54,000 SNPs versus human SNP chips with up to a million SNPs). Nonetheless, the data are expanding rapidly, driven by rapid advances in new sequencing technologies and genomics platforms. A 600K bovine SNP chip is expected later this year.

Although limited, the current datasets are already enabling initial genome-wide studies of natural selection in different species and populations. The genome can be surveyed without prior assumptions or biological hypotheses, resulting in a less biased set of loci which have putatively been subject to selection.

Also, genome-wide studies, in principle, provide a framework to distinguish between natural selection and population demographic history. The principal rationale/ assumption behind this is that population demographic history affects all loci equally, whereas positive selection acts on specific loci. Therefore, by sampling a large number of loci across the entire genome, appropriate statistics can be generated to quantify the genome-wide genetic variation and outliers can be identified as putative candidates of selection. Until now, the criteria in defining the outliers were relatively arbitrary (e.g. choosing a percentage cut off point). Other methods to better redefine the outliers are being investigated, including simulation studies and the use of models

[1] LPL Lau, DJ Lynn, DG Bradley are co-authors and participants in the Bovine Genome Sequencing and Analysis Consortium.

[2] LPL Lau, DJ Lynn, DG Bradley are co-authors and participants in the Bovine HapMap Consortium.

that integrate everything from population demographic history to mutational events that shape the genome.

1.4.2 The Human HapMap Project

Humans are by far the most studied species at a genome-wide scale. In 2003, The International HapMap Consortium published their first concerted effort to identify and catalogue genome-wide genetic variation in different human populations (Gibbs et al. 2005).

In this Phase I study, over a million SNPs were genotyped in 269 individuals (90 Yoruba in Ibidan, Nigeria – YRI; 90 in Utah, USA – CEU; 45 Han Chinese in Beijing, China – CHB; and 44 Japanese in Tokyo, Japan – JTP). As Han Chinese in Beijing, China and Japanese in Tokyo, Japan have similar allele frequencies, some analyses combined these two groups together as CHB+JPT. The SNPs for this study were evenly distributed across the autosomal chromosomes, with an average of 5kb inter-SNP space.

Using mainly the long haplotype tests, a variety of genes potentially subject to positive selection were detected. Some of these candidate loci included genes previously reported as being subject to selection (e.g. LCT for lactose tolerance and the Duffy gene (FY), which is responsible for resistance to malaria), but many of the strongest signals were in unexpected genes, which had not previously been hypothesised to be under selection (Gibbs et al. 2005) including SLC24A5 which has since been shown to influence skin pigmentation (Lamason et al. 2005).

It must be noted, however, that the SNP discovery process for SNPs used in the Phase I project was carried out on a small initial panel involving only a few individuals, resulting in an ascertainment bias being introduced into the dataset. As the probability of a particular SNP being included in the panel was a reflection of the allele frequency, common alleles were subsequently over-represented and rare alleles were more likely to be undetected. The allele frequency spectrum obtained was therefore distorted and effected any of the statistical tests that made use of the site frequency spectrum, including F_{ST}, nucleotide diversity and Tajima's D (Clark et al. 2005).

The Phase II HapMap study saw approximately another 2 million SNPs added to the number of SNPs genotyped and characterised, with an average inter-SNP distance of less than 1kb (Frazer et al. 2007a). Using long haplotype tests of LRH (Sabeti et al. 2002), iHS (Voight et al. 2006)), as well as a newly developed Cross Population Extended Haplotype Heterozygosity (XP-EHH) test (Sabeti et al. 2007), genes previously identified as being subject to selection (e.g. LCT, SLC24A5) again emerged as strong candidates of selection. Another finding of the analyses is the signature of selection on the LARGE gene in West African populations. The LARGE protein post-translationally glycosylates α-dystroglycan, which is a receptor for Lassa fever virus, and the site of the modification is critical for virus binding (Sabeti et al. 2007).

The Phase III HapMap analyses are currently underway. This phase involves genotyping of approximately 1.5 million SNPs in 11 populations, including the original 4 used in Phase I and Phase II studies.

1.4.3 Other HapMap Projects

Following in the footsteps of the International Human HapMap Consortium, several other studies have been established to catalogue and investigate genetic variation in other species, usually in tandem with genome sequencing projects of the corresponding species.

Over 8 million SNPs, for example, were identified by Frazer *et al.* in inbred mouse strains. This dataset was used to characterise the SNP distribution and the locations of both high and low SNP density regions, and to discern the ancestry of the haplotypes found in classical mice strains (Frazer et al. 2007b).

A rat SNP map consisting of about 3 million SNPs was generated and mapped to the draft genome sequence at a density of about one SNP per 800 base pairs. Just over 20,000 of these SNPs were selected for genotyping in 167 inbred strains and 64 recombinant inbred lines. The group identified 56 SNPs which are likely to affect protein function, of which seven of these are annotated to be involved in hereditary diseases or cancer including the alcohol dehydrogenase 2 gene (ALDH2) that is involved in acute alcohol intolerance and the proto-oncogene tyrosine kinase receptor ret precursor (RET) (Saar et al. 2008).

For the domestic dog, a SNP map consisting of over 2.5 million SNPs was generated and mapped to the draft genome sequence. The inter-SNP distance is approximately 1 per kb. While the SNP dataset was used in analyses to investigate linkage disequilibrium, haplotype structure and long-range haplotypes, the data were not analysed for evidence of selection between the different dog breeds. A 27K canine SNP array was also separately genotyped in an effort to map various canine Mendelian traits (Lindblad-Toh et al. 2005), including melanocyte-specific promoter of MITF which affects pigmentation in dogs (Karlsson et al. 2007).

A macaque SNP resource/database is also available. However, this dataset has not yet been comprehensively analysed (Malhi et al. 2007).

More recently, with the availability of bovine SNP genotyping panels, various analyses have been conducted including the search for strong candidates of positive selection in cattle populations. This forms part of the work of this thesis, of which some has been contributed towards the publication of the Bovine HapMap Consortium (Gibbs et al. 2009).

1.4.4 Genome Wide Bovine SNP Analyses

Genome wide SNP analyses are carried out on SNP chips, based on the principles of DNA microarray. The basic premise of SNP chip is that it contains target sequences attached to probes, with which when samples are tested, hybridisation of the samples and probes would emit signals which can then be recorded and interpreted into meaningful information.

Bovine SNP chips have been developed by member labs of the Bovine HapMap Consortium, subsequently improved, and the current commercial bovine SNP chip, BovineSNP50 by Illumina, contains an array of approximately 54,000 SNPs.

High density SNP assays for cattles have been developed based on the concept of deep sequencing to reduce the representation libraries, in order for the development process to be efficient and cost-effective. Pooled DNA from samples of populations of interest were used to construct the SNP libraries, and the sequencing effort was carried with the following ojectives in mind: (i) large target number of SNPs (>50,000), (ii) good level of coverage (1 SNP per 500bp), (iii) sequencing capacity based on budget, and (iv) the number of independent chromosomes sampled per reduced representation library (RRL) (Van Tassell et al. 2008).

In constructing the BovineSNP50 SNP assay, apart from the RRL, additional sources of SNPs used were from (i) markers derived from the bovine genome sequence assembly of a Hereford cattle, (ii) comparison of random shotgun reads from six cattle breeds to the Hereford genome assembly, (iii) alignments of sequence traces from NCBI of Holstein BAC (bacterial artificial chromosome) to the Hereford genome assembly, and (iv) filtered draft SNPs (Matukumalli et al. 2009). The effort yields a genotyping assay containing 54,001 SNPs with median interval of 37kb and maximum predicted gap of <350kb, with average minor allele frequency in the range of 0.24 to 0.27. However, it must be noted that the SNP discovery process has been biased towards taurine cattle breeds, thus alleles that are "indicine" in nature may have been missed and/or under-represented in the SNP assay.

1.5 Cattle and Its Domestication

Domesticated cattle are animals of significant economic importance, and according to a FAS/USDA report produced in November 2007, the cattle population worldwide has an estimated 996 million head, (http://www.fas.usda.gov/dlp/circular/2007/livestock_poultry_11-2007.pdf). Cattle have been domesticated for various purposes, including as a food source (meat), for milk and other dairy production, for agricultural use (draught ability), and for leather goods production (hides to make shoes and clothing). The importance of cattle had been indicated for millennia, with earliest manifestations in the form of cave paintings.

1.5.1 The Origin and Domestication of Cattle

Modern cattle are typically classified as either taurine (*Bos taurus*) or zebu (*Bos indicus*) cattle. Cattle domestication took place approximately 10,000 years ago from the varieties of the wild ancestor aurochs (*Bos primigenius*) (Helmer et al. 2005). Domesticated cattle and the aurochs co-existed for thousands of years, until the extinction of the aurochs following the death of the last auroch in Europe in 1627. The aurochs once ranged across Europe, Northern Africa and Southern Asia, where they existed as regional subspecies; *Bos primigenius primigenius* (European aurochs), *Bos primigenius opisthonomus* (African aurochs) and *Bos primigenius namadicus* (Asian aurochs).

The Fertile Crescent in the Near East, ranging from the Persian Gulf in the East to the Mediterranean Sea in the West, has been found to be the domestication centre of cattle (Helmer et al. 2005). One hypothesis was that a single domestication event took place, and subsequently as humans migrated, the expansion of cattle occurred northward into Europe, westward into Africa and eastward into Asia.

Analyses of mtDNA revealed, however, a pronounced dichotomy in the phylogenetic tree, where European and African *Bos taurus* formed one clade, and *Bos indicus* formed another. The two clades are highly divergent, with a divergence time estimated to be over 100,000 years ago (Loftus et al. 1994). As this divergence time is long before the time of domestication, speculation arose that at least two domestication events had taken place – the domestication of European *Bos primigenius primigenius* to give rise to the modern *Bos taurus*, and domestication of Asian *Bos primigenius namadicus* to give rise to modern *Bos indicus*.

A further third centre of domestication has been suggested in Northern Africa, where it has been suggested that African *Bos taurus* was domesticated from *Bos primigenius opisthonomus*. Data from mtDNA has estimated the divergence time between European and African *Bos taurus* as around 22,000 years ago, which is longer than predicted (Bradley et al. 1996). This hypothesis is, however, controversial. Beja-Pereira et al postulated that a more parsimonious explanation is a simple founder effect that would have given rise to a high frequency common mitochondrial haplotype, which is observed in Northern African populations but is at very low frequency in the Near East and completely absent from most European populations. This could have occurred following the migration of a small number of animals into Northern Africa from the Near East (Beja-Pereira et al. 2006).

1.5.2 Modern Cattle

Modern cattle are categorised as either taurine or indicine cattle. Taurine cattle are of European or (West) African origins whereas the indicine cattle, or zebu, originated on the Indian subcontinent. The zebus are better adapted to arid regions, and are characterised by a hump and dewlap, and have larger sweat glands to cope with the tropical environments. There are over 1,000 cattle breed described by Felius in "Cattle Breeds: An Encyclopedia", of which about 75 are zebus. Many of the contemporary breeds are the result of cross-breeding between two or more of the older breeds. Sanga cattle in Africa, for example, are crossbreeds between zebu and indigenous humpless cattle, and they have smaller humps in comparison to the pure zebu.

A range of cattle breeds have been used in the work discussed in this thesis, and were used to represent cattle populations from distinct geographical groups. [3]

1.5.2.1 European Cattle Breeds

European *Bos taurus* is represented by a large number of breeds. Several were sampled for work in this thesis. Friesian cattle are large dairy cattle originating from Friesland in the province of North Holland. Characteristically black and white pied coat in appearance, Friesian cattle were exported to America until disease problems led to a cessation of movement. Over time the regionalisation meant the two cattle groups became quite different from their counterpart and for this reason, the American stock is now known as Holstein while Friesian refers to those of traditional European ancestry. Crosses between the two are known as Holstein-Friesian.

Aberdeen Angus are beef cattle that originated in Aberdeenshire in Scotland; they are usually solid black in coat colour, and are resistant to harsh weather while remaining good natured and undemanding. A closely related breed, Red Angus, have a red colour coat. The Angus is polled, i.e. naturally without horn. Hereford is a hardy breed from Herefordshire, England. They are bred for beef in temperate areas, and most have short thick horns that curve down at the side of the head. Charolais is also a breed of beef cattle, originating from Central France. Their coat is almost pure white, and the meat contains little fat. Another French beef cattle breed is Limousin, which produces lean meat and these animals are very easy to manage, with little calving problems. Romagnola is an Italian beef cattle breed that is robust and muscular.

Guernsey and Jersey are both dairy cattle from the British Channel Islands of the same names. Brown Swiss is also a dairy cattle breed, originating from the Alps in Switzerland. Bred on a harsher condition, Brown Swiss are resistant to heat, cold, and can subsist with little care or feed.

[3] Descriptions of cattle breeds were obtained from "Cattle Breeds: An Encyclopedia" published by Felius in 1995 and University of Oklahoma - Cattle Breeds website http://www.ansi.okstate.edu/breeds/cattle/

Norwegian Red is a relatively recent breed, having been bred for superior dual-purpose characteristics, with its gene pool heterogeneously contributed to by other cattle breeds including Ayrshire, Friesian and Holstein. Piedmontese is also a dual-purpose cattle breed, but originating from Italy. The animals are highly muscular due to mutation in myostatin, which also reduces fat content while improving meat quality.

1.5.2.2 African Cattle Breeds

African breeds that are relatively pure (i.e. with low levels of *Bos indicus* introgression) were used in this thesis to represent African *Bos taurus*. N'Dama is a trypanotolerant, longhorn breed found all across Western Africa despite endemic trypanosomiasis (also known as sleeping sickness). Somba cattle are also trypanotolerant, and this small West African shorthorn breed originated from the Atacora highlands of Northern Benin. Lagune is the smallest of the West African shorthorns, although morphologically not unlike the Somba.

1.5.2.3 Indicine Cattle Breeds

Bos indicus populations investigated in this thesis are represented by several breeds. Hariana, Tharparkar and Sahiwal originated from the Northern part of the Indian subcontinent. Hariana cattle are shorthorns bred for draught ability and dairy production. Tharparkar cattle are lyrehorns, also bred for draught ability and dairy production, but require constant human contact or are apt to be wild and vicious. Sahiwal originated from the Punjab region and it is one of the best dairy breeds on the subcontinent. It is also tick-tolerant, heat-tolerant and has high resistance to parasites.

From the Southern Indian subcontinent, the zebus are represented by Ongole, Nelore, Gir and Brahman. Ongole cattle are large, shorthorn docile animals bred for draught ability and dairy production. Nelore is a Brazilian breed derived from Ongole, prized for its fast growing rate and parasite resistance. Unlike Ongole, Nelore is bred for beef. Gir is a dairy breed originated from the Gujarat state and was used in the development of

Brahman in North America. Brahman is a hardy breed of cattle which is considered sacred in India. American Brahman, however, is a beef cattle breed which was developed from four different Indian cattle breeds – Gir, Nelore, Guzerat and the Krishna Valley.

1.5.2.4 Cross-Breed Cattle

Several *Bos taurus/Bos indicus* crossbred cattle were also represented. Beefmaster and Santa Gertrudis are two beef cattle which were developed in Texas, America. Beefmaster was developed in early 20th century from cross-breeding Hereford and Shorthorn[4] cattle to American Brahman bulls principally of Gir breeding. Modern estimates of the composition of Beefmaster is just under half Brahman, just over a quarter Hereford and just over a quarter Shorthorn. Santa Gertrudis is a crossbreed of Brahman bulls and Shorthorn cows.

Sheko is a breed of shorthorn, humpless cattle and is believed to be the last remnants of indigenous humpless cattle in East Africa. Modern Sheko cattle are mainly interbred Sheko and Sanga, and recent phylogenetic study showed very high frequency of indicine allele frequency (90%) and low taurine allele frequency (10%) in Sheko males.

1.5.2.5 Outgroup Animals

Non-cattle outgroups have also been used in the work of this thesis to facilitate the determination of ancestral and derived alleles at loci of interest. Water buffalo (*Bubalus bubalis*) is a multi-purpose Asian livestock. Anoa is a subgenus of buffalo which is native to Indonesia, and is essentially a miniature water buffalo. Gaur (*Bos gaurus*) is a wild bovid found in South Asia and Southeast Asia. Yak (*Bos grunniens*) is a long-haired bovine found in South-central Asia, the Tibetan Plateau and Mongolia.

[4] Shorthorn breed of cattle originated from north-eastern coast of England. It was developed as dual purpose breed but over time has diverged. In the second half of 20th century, two separate breeds had developed, one for beef and one for dairy.

Plains bison (*Bison bison bison*) are nomadic grazers from North America. Cattle can interbreed with the closely related gaur, yak and bison. They cannot, however, successfully interbreed with the more distantly related buffalo.

1.5.3 Domesticated Cattle and Disease Challenges

Cattle has been domesticated and selectively bred for desirable traits including beef quality, milk yield, draught ability, temperament, growth rate and coat colour. The process of domestication has also led them to closer proximity to human and other domesticated species. These changes are significant to both the environment and the population structure of cattle, and exposed the cattle to new disease challenges.

In different breeds and populations, different innate resistances are exhibited. Several of the West African breeds such as N'Dama and Somba are trypanotolerant, whereas the zebus are more tick-tolerant and parasite resistant. With co-existence in close quarters with human and other livestock, close-species transmission would have introduced novel diseased into the population. Phylogenetic analyses of human pathogens have shown that a number of human diseases, such as measles (rinderpest) and pertussis, may have arisen as a result of domestication. On the other hand, such analysis has also indicated that while domestication was not the underlying cause of the presence of pathogens such as tapeworms in human, it has enabled the opportunity for human tapeworms to infect livestock such as cattle and pigs (Pearce-Duvet 2006).

An increase in population size also affects the fitness of the herd. Certain pathogens, such as *Mycobacterium bovis* which causes bovine tuberculosis, predate domestication but within confined area and higher population density, the diseases have a great opportunity to spread given the availability of hosts, and may infect other animals faster than what would otherwise be observed in the wild.

One way to identify loci that are most significant in resisting diseases encountered following domestication is through direct evidence that they have undergone positive selection in recent bovine history. Similarly, other traits which underlying genes have been subject to selective pressure would have left their signatures in the bovine genome. Identifying these loci is an important step in better understanding of genotypic evolution given phenotypic changes observed, and in terms of immune-related genes, this may provide vital clues in combating disease challenges faced by modern cattle.

1.6 Thesis Structure

This thesis will present work that has been carried out to detect positive selection in mammalian genomes, with particular interest in the bovine genome. Chapter 2 of the thesis introduces a strategy to detect positive selection in the extracellular domain of human and chimpanzee, with a view to extend the method for interspecies detection of positive selection in cattle in Chapter 3. Collaborative work with the Bovine Genome Consortium will also be discussed in this chapter. Chapter 4 explores the idea of utilising expressed sequence tags (ESTs) to generate a SNP dataset and to elucidate the effect of polymorphisms that are tolerant or intolerant, as well as dramatic stop codon polymorphisms in the genome. Analyses to identify candidates of positive selection using genome-wide SNP datasets, including those contributed towards Bovine HapMap Consortium, makes up Chapter 5 of this thesis. Chapter 6 concludes this thesis.

2. Detection of Positive Selection in Human and Chimpanzee Extracellular Domain

2.1 Positive Selection in Human and Chimpanzee

Human and chimpanzee shared a common ancestor as recent as 4-5 million years ago (Stauffer et al. 2001; Hobolth et al. 2007) but the phenotypic divergence of human from the great apes following the speciation has been remarkable. We acquired much larger brains, became bipedal, and developed speech, among many other traits. We are also karyotypically different, with one less chromosome than the chimpanzee (the counterpart of human chromosome 2 is chimpanzee chromosomes 2a and 2b (formerly chromosomes 12 and 13) and a few other discreet differences, but, otherwise, the broad-scale organisation of the two genomes remains highly conserved (Yunis, Sawyer, and Dunham 1980; Yunis and Prakash 1982; Gagneux and Varki 2001). The two genomes share over 98% of nucleotide identity (Li and Saunders 2005).

It has been proposed that genes which have undergone positive selection specifically on the human lineage may be the genes that make us different from our closest evolutionary relative, the chimpanzee. The availability of genome sequences from various species has now enabled a systematic approach in comparative genome studies to shed some light on this hypothesis. Several pivotal studies involving comparative studies of human and chimpanzee genomes have been published in recent years.

2.1.1 Clark et al. – A study of Human-Chimpanzee-Mouse Orthologous Gene Trios

Clark *et al.* (Clark et al. 2003) in 2003 compared a set of 7,645 orthologous human, chimpanzee and mouse genes to investigate patterns of divergence between the human and chimpanzee lineages. Two statistical tests were carried out to identify genes which have undergone positive selection.

In their first test, they tested the neutral theory null hypothesis by fixing $d_N/d_S = 1$ in the human lineage. The alternative hypothesis allowed for d_N/d_S to vary across all 3 branches. In the second test, they used the branch-site model (Yang and Nielsen 2002) which allows the d_N/d_S ratio to vary among the sites within each lineage and among the lineages under investigation. Their null hypothesis assumed all sites evolved either neutrally ($d_N/d_S = 1$) or under purifying selection ($d_N/d_S < 1$) in the human lineage, while in their alternative hypothesis, some sites evolved with an accelerated amino acid substitution ($d_N > d_S$). A likelihood ratio test (LRT) was then conducted to determine which of the models was statistically favoured. This was similarly done for the chimpanzee lineage.

Clark *et al.* found a total of 1,547 human genes and 1,534 chimpanzee genes that met the criteria for positive selection, i.e. had $d_N/d_S > 1$. However, in the first test of the neutral null hypothesis, only 6 genes with $d_N/d_S > 1$ had statistical support ($P < 0.05$). In the second test, they found 125 genes with $d_N/d_S > 1$ ($P < 0.01$). They attributed the differences to the ability of the second test to predict genes of which overall d_N/d_S may be low but contained domain(s) that had undergone positive selection.

Clark *et al.* used the results from the second test for further analysis, including categorising the genes based on PANTHER ontology (Thomas et al. 2003a; Thomas et al. 2003b; Mi et al. 2005) (see section 2.2.6 for further explanation), as well as correlating these genes to Mendelian disorders. They found genes involved in olfaction and amino acid catabolism showed human-specific acceleration in protein evolution and postulated that differences in the lifestyles between human and chimpanzee may have lead to alternative selective pressure on these genes. Other gene categories found to be under selection in this study included human developmental genes, and speech and hearing genes.

It would have seemed like the implementation of the branch-site model indeed uncovered numerous candidate genes under adaptive evolution. However, as the second test was implemented based on the assumption that the relative branch lengths leading to human, chimpanzee and mouse are the same for all genes, relaxation of selective constraint may have given rise to an excessively long human branch and the genes may have been incorrectly identified to be under positive selection (Eyre-Walker 2006).

Moreover, a later computer simulation study by Zhang (Zhang 2004) revealed that the detection of positive selection by the branch-site method alone led to numerous cases of false positive prediction of positive selection. Simulations showed an unacceptably high (20%-70%) false positive rate, prompting calls for re-evaluation of past results obtained primarily from implementing the branch-site likelihood method. Zhang *et al.* (Zhang, Nielsen, and Yang 2005) have since modified and improved the branch-site model which is now reported to be more robust and to have lower false positive rates.

2.1.2 Nielsen et al. – Pairwise Analysis of Human and Chimpanzee Genes

Nielsen *et al.* (Nielsen et al. 2005) reported an analysis of 20,361 human and chimpanzee genes in a study in 2005. The 7,645 genes analysed by Clark *et al.* (Clark et al. 2003) were also included in this dataset. Nielsen *et al.* first eliminated 6,630 predicted genes without a hit to known genes in public databases, thus reducing the dataset to 13,731 genes. They also further reduced the gene number to 8,079, excluding genes with fewer than three mutations, genes shorter than 50 base pairs in length, and genes which contained internal stop codons. Within this conservative 8,079 gene dataset, 3,913 genes had previously been analysed by Clark *et al.* (Clark et al. 2003).

Nielsen *et al.* found 733 genes with $d_N/d_S > 1$ from this pairwise analysis. Amongst them, only 35 genes had $P < 0.05$, determined from the likelihood ratio test carried out and compared to a simulated distribution of the test statistics. The usual χ^2 distribution was not used as the level of divergence between human and chimpanzee pairwise sequences is very low and thus the assumptions of the χ^2 distribution are violated.

Nielsen *et al.* also mapped the genes to PANTHER ontology terms and carried out further analysis on the top 50 genes with strongest evidence of selection including genes involved in immunity and defence, spermatogenesis and cancer. The pairwise method, which averages d_N/d_S over all the codon sites in the alignment, is limited as it cannot distinguish which of the two lineages i.e. human or chimpanzee, or indeed whether both human and chimpanzee, were under positive selection.

2.1.3 The CSAC – Initial Comparison of Chimpanzee to Human

Later in the same year, the Chimpanzee Sequencing and Analysis Consortium (CSAC) (CSAC 2005) published the results from initial comparative study of the chimpanzee genome to the human genome. In order to elucidate the nature of chimpanzee gene evolution, a set of 13,454 orthologous human-chimpanzee gene pairs was identified. As a chimpanzee gene catalogue was not available, the CSAC aligned human genes to the chimpanzee genome through the application of BlastZ (Altschul et al. 1990) and identified the orthologous chimpanzee bases.

The human and the chimpanzee genomes, each containing approximately 3 billion nucleotides, share >98% nucleotide identity. This means, however, that there are an estimated 35 million nucleotide differences between them (Li and Saunders 2005). Nonetheless, when the orthologous genes between these species were examined, about 29% of them were determined to be identical and typically differed by only 2 amino acids, one on each lineage.

Since there were few synonymous changes in a typical gene, and often none, the chimpanzee genome sequence was used to generate an estimation of the local intergenic/intronic substitution rate where appropriate. Given as K_I, it was used to calculate K_A/K_I ratio, where K_A was the number of nonsynonymous substitutions per nonsynonymous site and K_A/K_I has the same implication as d_N/d_S. Both ratios can be denoted using ω.

The CSAC found a total of 585 genes with $K_A/K_I > 1$, which was more than twice as many as expected (263) to occur simply by chance, given a simulation study that allowed purifying selection to act in a non-uniform manner across the genes. Some of the most extreme outliers include genes which are involved in immunity, for example glycophorin C which mediates one of invasion pathways in human erythrocytes by *Plasmodium falciparum*, and granulysin which mounts antimicrobial activity against intracellular pathogens such as *Mycobacterium tuberculosis*. As these genes were found using pairwise analysis, it was not possible to differentiate on which lineage selection had taken place.

The CSAC also identified a set of unambiguous 7,043 orthologous human-chimpanzee-mouse-rat gene quartets to conduct a comparison of the hominid (i.e. human and chimpanzee) and murid (i.e. mouse and rat) lineages. It is also of note that genes belonging to large gene families were omitted in the analyses, due to difficulty of 1:1:1:1 orthology assignment across hominids and murids. One of the largest families of genes that is known to undergo such rapid divergence are the olfactory receptors.

In addition, the CSAC mapped the genes studied to the Gene Ontology (GO) classifications (Ashburner et al. 2000) (see section 2.2.6 for further explanation). The GO categories which had elevated ω_{hominid} included immunity and host defence, reproduction, olfaction and apoptosis while the GO categories with low ω_{hominid} included intracellular signalling, metabolism, neurogenesis and synaptic transmission.

The CSAC commented that the gene categories found by Clark *et al.* (Clark et al. 2003) to be under positive selection have been annotated as gene categories which showed accelerated divergence. This led to the CSAC to speculate that the results by Clark *et al.* may be enriched for false positives in categories that underwent strong relaxation of constraints in the hominids, or there may be some correlation between positive selection and relaxation of constraints.

2.1.4 Arbiza *et al.* – Positive Selection, Relaxation and Acceleration

Following the publication of the improved branch-site likelihood model (Zhang, Nielsen, and Yang 2005), Arbiza *et al.* undertook a study to identify genes which have been positively selected in the common ancestral lineage of human and chimpanzee, using mouse, rat and dog as weighted outgroups (Arbiza, Dopazo, and Dopazo 2006). This method was used to differentiate events of relaxed selective constraints from accelerated evolution due to positive selection.

A total of 14,185 orthologues were identified, of which 4,511 were subsequently excluded from further analyses following removal of sequences with fewer than 3 unique base pair differences. The remaining 9,674 genes were analysed using Test 1 (i.e. test of either relaxation of constraint or positive selection on foreground branch) and Test 2 (i.e. direct test of positive selection on foreground lineages) of the branch-site model as described by Zhang *et al.* (Zhang, Nielsen, and Yang 2005) on both the human and the chimpanzee lineages. Arbiza *et al.* found, with correction for multiple testing, a total of 108 human and 577 chimpanzee genes to be true cases positive selection and not falsely identified due to relaxation of selective constraints.

Arbiza *et al.* mapped these genes to GO terms to elucidate the associated functions of these genes. They found genes involved in cellular protein metabolism, G-protein coupled receptor signalling, sensory perception, transcription and transcription regulation, and immune response, among others. They also noted that only a small number of genes are common between human and chimpanzee, showing that the evolution of genes through adaptive evolution occurs frequently after speciation, and not at the common ancestor level.

2.1.5 Positive Selection in Extracellular Domains

The relatively small number of genes with strong evidence for positive selection emerging from these studies led to speculation that human-chimpanzee divergence was perhaps not really due to adaptive protein coding changes, but rather more related to adaptive changes in gene expression (Nielsen 2006). Undoubtedly there have been adaptive changes in gene expression (Gilad et al. 2006) but it is also likely that the relatively conservative tests applied in the studies discussed makes it difficult to detect cases of positive selection on protein-coding changes.

Indeed, all of the tests applied, with the exception of the branch-site model, are conservative in that they require d_N/d_S to be greater than one over entire gene. It is more usual for positive selection to act on particular codons or domains in the gene over a short period of evolutionary time while purifying selection acts on the majority of other sites (Zhang 2004). Because of this, signals of positive selection may be swamped in this background of purifying selection. We postulated that if one could identify the regions where selection was likely to be strongest these regions could be isolated and screened for signatures of positive selection that would be missed in whole gene sequence analyses.

The extracellular domains of proteins are the components that interact directly with the external environment and are involved in a range of interactions including ligand and antigen binding, and are frequently points of contact for pathogen interactions. One example is the interaction between HIV and one of its host receptors, CCR5. The extracellular domain of CCR5, which is exploited by HIV to gain entry into cells (Lusso 2006), has been shown to undergo strong selection in human populations (Hedrick and Verrelli 2006).

Cluster of differentiation 2 (CD2) is a cell-surface protein found on T-cells and natural killer cells and is implicated in mammalian defence (Davis et al. 2003). Lynn *et al.* (Lynn et al. 2005) also demonstrated that the power of detecting for positive selection of CD2 was increased when only the extracellular domain was tested.

This led to the proposal that extracellular domains may be strong candidates to examine for signatures of adaptive evolution in genes, which could have been missed in the whole gene sequence analysis. In this study, regions of human and chimpanzee genes encoding the extracellular domains were systematically analysed for evidence of positive selection.

2.2 Materials and Methods

2.2.1 Human-Chimpanzee-Mouse Orthologues

Human, chimpanzee and mouse (HCM) gene entries were downloaded from the NCBI GenBank database (ftp://ftp.ncbi.nih.gov/genbank/) (Benson et al. 2006). A total of 29,549 human, 21,795 chimpanzee and 56,815 mouse Reference Sequence (RefSeq) genes were extracted using CODERET (Mullan and Bleasby 2002). The number of mouse genes was nearly twice as many as the others due to the presence of multiple transcripts of some RefSeqs in the dataset downloaded.

An all-against-all BlastP (Altschul et al. 1990) was carried out against the protein dataset from the three species with an E-value cut off of e^{-10} and a Perl script was written to extract the best blast hits. In order to deal with cases with multiple transcripts, the script was modified to include GeneID information. When more than one hit of the same GeneID were found, the program would treat these as one gene until a different GeneID was encountered, i.e. a different gene. The extracted blast hits were listed and ranked by their E-values. The best blast hit was assigned when the second best hit was at least e^{10} worse than the top hit. In cases where multiple gene transcripts were present, the transcript version that has the best score within the group is identified as the best blast hit.

The three lists of best blast hits (one each for human, chimpanzee and mouse) were loaded into MySQL database (http://www.mysql.com/) and subsequently queried to obtain 3-way reciprocal best hits.

2.2.2 Assessment of Orthologous Gene Dataset

In order to confirm the accuracy of orthology prediction, 1:1:1 human, chimpanzee and mouse orthologues were downloaded from NCBI's HomoloGene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene) which is a database of homologues detected through an automated system based on annotated genes of completely sequenced eukaryotic genomes, and compared to our orthology predictions. Our dataset was found to be in 93% agreement with HomoloGene.

2.2.3 AHMM Prediction of Extracellular Domains

TM-HMM is one of the most widely used transmembrane protein structure prediction programs (Krogh et al. 2001; Moller, Croning, and Apweiler 2001) (http://www.cbs.dtu.dk/services/TMHMM-2.0/). It identifies the locations of the transmembrane helices as well as the intervening loop regions of given proteins. Using this information, the location of the extracellular domain can be inferred. TMHMM ranked very well in comparison to other transmembrane domain prediction algorithms such as HMMTOP, TMPred and MEMSAT (Lao, Okuno, and Shimizu 2002; Melen, Krogh, and von Heijne 2003). However, for single-pass transmembrane proteins, the prediction of the intracellular and extracellular topologies may be reversed and therefore inaccurate.

AHMM, an extension of TM-HMM, was recently developed by Xu *et al.* (Xu, Kearney, and Brown 2006). AHMM aims to improve the prediction accuracy of TM-HMM through incorporation of functional information from Prosite (Hulo et al. 2006). Regions of the protein that contain Prosite domains that are commonly associated with extracellular functions are assigned higher probabilities of being located outside the cell. Similarly, for domains that are associated with intracellular functions, these regions have increased intracellular scores. The inclusion of Prosite functional domain information in AHMM improves the accuracy of TM-HMM by 39%.

AHMM was applied to the human protein sequences and the regions encoding the extracellular domains were extracted from proteins containing at least one transmembrane pass. Extracted extracellular domain regions that were less than 20 amino acids in length were excluded.

Full length human, chimpanzee and mouse orthologous proteins were aligned using T-Coffee (Notredame, Higgins, and Heringa 2000). Treating these alignments as profiles, a second run of T-Coffee was implemented to append the extracellular domain predictions to the first run alignments. A Perl script was used to extract the sequences corresponding to the human prediction in all the three species. A second Perl script was applied to align the coding sequences based on the extracted protein alignments, which acted as templates, so that appropriate gaps were included. Following gapped coding sequence alignments, another 10 genes were excluded from the subsequent PAML analysis due to the presence of missing or gapped data in this region in at least one of the orthologous genes. Figure 2.1 illustrates the method used in extracting the coding sequence alignments of the extracellular domains of human, chimpanzee and mouse orthologous trios.

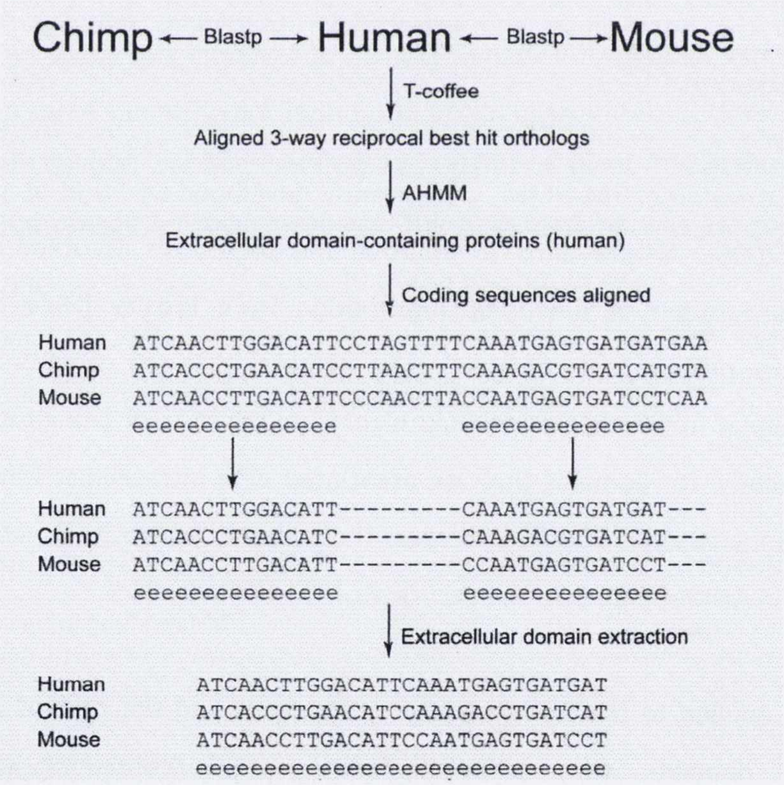

Chimp ← Blastp → Human ← Blastp → Mouse
T-coffee
Aligned 3-way reciprocal best hit orthologs
AHMM
Extracellular domain-containing proteins (human)
Coding sequences aligned
Human ATCAACTTGGACATTCCTAGTTTTCAAATGAGTGATGATGAA
Chimp ATCACCCTGAACATCCTTAACTTTCAAAGACGTGATCATGTA
Mouse ATCAACCTTGACATTCCCAACTTACCAATGAGTGATCCTCAA
eeeeeeeeeeeeeee eeeeeeeeeeeeeee
Human ATCAACTTGGACATT---------CAAATGAGTGATGAT---
Chimp ATCACCCTGAACATC---------CAAAGACGTGATCAT---
Mouse ATCAACCTTGACATT---------CCAATGAGTGATCCT---
eeeeeeeeeeeeeee eeeeeeeeeeeeeee
Extracellular domain extraction
Human ATCAACTTGGACATTCAAATGAGTGATGAT
Chimp ATCACCCTGAACATCCAAAGACCTGATCAT
Mouse ATCAACCTTGACATTCCAATGAGTGATCCT
eeeeeeeeeeeeeeeeeeeeeeeeeeeeee

Figure 2.1 | Illustration of the method used to extract the coding sequence alignments of human, chimpanzee and mouse orthologous extracellular domains.

2.2.4 CODEML Analysis of Extracellular Domains

CODEML from the PAML 3.14 suite of programs (Yang 1997; Yang 2007) was implemented on our dataset which contained 1,948 AHMM-predicted extracellular domains of orthologous HCM genes. CODEML is a method frequently used to fit models of evolution by maximum likelihood to the alignment of protein coding sequences in order to estimate the d_N/d_S ratio.

Two models were implemented to test the statistical significance of the selective pressures on each lineage. In the one-ratio model (NSsites = 0, model = 0) each lineage was modelled to have the same d_N/d_S ratio. The ratio is constrained between 0 and 1, and thus does not allow for the presence of positive selection. The second model, the free-ratios model (NSsites = 0, model = 1), is a model of selection which estimates an independent d_N/d_S ratio for each of the three lineages. The value of d_N/d_S in the free-ratios model may exceed 1.

The two models were compared by Likelihood Ratio Test (LRT), calculated from the log likelihood (lnL) values of both models. Twice the difference between $\ln L_{\text{free-ratios}}$ and $\ln L_{\text{one-ratio}}$ (i.e. $2\Delta\ell$) was compared to a χ^2 distribution in order to obtain the *P* values. The degree of freedom for the distribution was calculated from the difference in the number of parameters between the two models tested. The *P* values were corrected using the Benjamini and Hochberg false discovery rate correction for multiple testing (Benjamini and Hochberg 1995a). Evidence of positive selection was inferred when d_N/d_S was greater than 1 on the human and/or the chimpanzee lineages, and the free-ratios model was significantly favoured over the one-ratio model.

A more stringent test of positive selection on the human lineage was then carried out on each of these genes, comparing the one-ratio model to a model where the human lineage was selected as the "foreground" lineage and its d_N/d_S was specifically allowed to vary (NSsites = 0, model = 2). Correction for multiple testing was also conducted for this analysis.

2.2.5 Analysis of Human and Chimpanzee Pairwise Alignments

Methods that use an outgroup species such as the method described above are useful, in that the outgroup allows the direction of selection (i.e. whether it acts on the human or chimpanzee lineage) to be inferred. On the other hand, the inclusion of an outgroup results in the preclusion of numerous genes that could be analysed between the two species of interest due to difficulties in orthology assignment. In an effort to maximise the number of genes analysed in this study, an implementation of a pairwise approach similar to that of Nielsen *et al.* (Nielsen et al. 2005) was carried out.

Human coding sequences corresponding to the extracellular domains were aligned to the chimpanzee genome (panTro2 downloaded from UCSC FTP site at ftp.hgdownload.cse.ucsc.edu) using Blast-like Alignment Tool, BLAT (Kent 2002). BLAT performed optimally in aligning the coding sequences of the genes to the chimpanzee genome for extraction and therefore the pairwise alignments between human and chimpanzee were of high quality. Predicted domains that were less than 60 base pairs in length, sequences that could not be uniquely aligned to the genome and sequences containing stop codons were all excluded from the analysis. The regions of the chimpanzee genome to which the human extracellular domains aligned were extracted. ClustalW (Thompson, Higgins, and Gibson 1994) was used to generate human-chimpanzee pairwise alignments.

Pairwise d_N/d_S ratios were estimated under two models of CODEML (Yang 1997; Yang 2007). In the first model, the d_N/d_S ratio was fixed to be equal to one, while the d_N/d_S ratio of the second model was estimated by maximum likelihood from the data. These two models were compared by means of LRT. The low divergence of human and chimpanzee pairwise sequences violates the assumptions of χ^2 distribution (Nielsen et al. 2005). We therefore simulated an empirical distribution of the LRT statistics, given the sequence divergence and the shorter alignment lengths. *P* values were calculated based on this simulated distribution.

2.2.6 Assignment of Ontologies: PANTHER Ontology and Gene Ontology (GO)

In order to classify the biological functions of genes with evidence of positive selection, each gene was mapped to two different ontology databases.

Protein ANalysis THrough Evolutionary Relationships, or PANTHER, is an ontology database available at http://www.pantherdb.org and contains sets of controlled vocabulary in terms of biological processes and molecular functions related to a large curated collection of protein families/subfamilies (Thomas et al. 2003a; Thomas et al. 2003b). In addition, the most recent PANTHER database also contains associated biological pathway information (Mi et al. 2005). These terms are matched based on the transcript/protein/gene ID.

Gene Ontology (GO) (Ashburner et al. 2000) is another ontology database, available from http://www.geneontology.org and is similar to PANTHER. It provides structured and controlled vocabularies and classifications used in the annotation of gene sequences. GO however, supplies slightly different categories of terms from PANTHER. The categories found in GO relate to biological processes, molecular functions and cellular components.

Subsequent development of GO produced "GO slims" (Harris et al. 2004), a high-level view of each of the three existing ontologies. This allows the grouping of terms into a broad range of high-level categories to provide a more focused view of the biological processes, molecular functions and cellular components involved in a dataset.

2.2.7 Human-Chimpanzee-Rhesus Macaque Orthologous Trios

Shortly following the completion of our work on this analysis, the sequence of the genome of rhesus macaque was published (Gibbs et al. 2007). It prompted a re-analysis of our orthologous trio method, substituting mouse sequences with macaque sequences as the outgroup. The human-macaque divergence time of approximately 25 million years (Gibbs et al. 2007) is better placed in comparison to human-mouse divergence time of approximately 75 million years (Waterston et al. 2002), making macaque a more suitable outgroup.

Due to poor gene annotation for the macaque, the human-macaque orthologous gene set was generated by using BLAT to align the coding sequences of the human genes to the macaque genome (rheMac2 also downloaded from UCSC FTP site). It resulted in a total of 24,985 human-macaque orthologous gene pairs.

The lists of human-chimpanzee and human-macaque orthologues were loaded into a MySQL database (http://www.mysql.com/) and subsequently queried to obtain 22,366 3-way reciprocal best hits. The sequences were extracted for extracellular domain and analysed using CODEML from PAML 4.0 (Yang 2007), as per protocol in Sections 2.2.3 and 2.2.4. In order to avoid confusion this dataset is denoted as the HCR (human, chimpanzee, rhesus macaque) dataset, and a total of 4,745 HCR trios which contain extracellular domain predictions were analysed.

2.3 Results

2.3.1 AHMM-Predicted HCM Extracellular Domain-Containing Proteins

The AHMM program predicted 6,609 human proteins as containing extracellular domain that was at least 20 amino acids in length. 1,948 of these had putatively 1:1:1 orthologues in chimpanzee and mouse. 592 human and 393 chimpanzee orthologues were found to have $d_N/d_S > 1$ in the region encoding the extracellular domain. Of these, 192 human and 113 chimpanzee genes were also significant under the LRT that compared the free-ratios model to the one-ratio model (corrected $P < 0.05$) (Supplementary Table 2.1). There were 34 genes putatively subject to positive selection in both species (Table 2.1).

As the human and the chimpanzee genomes are highly conserved (>98%) (CSAC 2005), often there are little or no synonymous changes between the primate gene pairs in comparison to their mouse orthologues. When the rate of synonymous changes (d_S) is very small, or indeed nil, it contributes to highly elevated and inaccurate d_N/d_S ratios. Chamary *et al.* (Chamary, Parmley, and Hurst 2006) in their recent review stated that strong purifying selection on synonymous sites could give rise to $d_N/d_S > 1$ as the d_S value is very low. Our dataset was therefore filtered to exclude genes with $d_S = 0$. This resulted in the reduction of the number of human candidate genes with $d_N/d_S >1$ from 592 to just 75, indicating the scale of this problem. Of these, 44 genes are also significant under the LRT comparing the free-ratios model to the one-ratio model (Figure 2.2) and only 9 are found in common with the genes list in Table 2.1. They are marked with light grey background in the table.

Table 2.1 | Genes with $d_N/d_S > 1$ on both the human and the chimpanzee lineages ($LRT_{free\text{-}ratio}$, $P < 0.05$). Genes that are also found following $d_S = 0$ filter step is shaded in light grey.

Human RefSeq	Chimp RefSeq	HUGO Symbol	Gene Description	Human d_N/d_S	Chimp d_N/d_S	$LRT_{free\text{-}ratio}$	$LRT_{one\text{-}ratio}$	$2\Delta\ell$	P value
NM_007073	XM_527578	BVES	blood vessel epicardial substance	1.6305	1.8705	-1467.6005	-1496.6155	58.0301	0.0000
NM_138966	XM_512172	NETO1	neuropilin (NRP) and tolloid (TLL)-like 1	999.0000	1.3653	-1867.2820	-1893.5189	52.4738	0.0000
NM_004733	XM_516831	SLC33A1	solute carrier family 33 (acetyl-CoA transporter), member 1	36.4591	2.1561	-621.5033	-645.2229	47.4392	0.0000
NM_182607	XM_521214	VSIG1	V-set and immunoglobulin domain containing 1	106.6730	1.9474	-1685.9368	-1707.5292	43.1849	0.0000
NM_052944	XM_510886	SLC5A11	solute carrier family 5 (sodium/glucose cotransporter), member 11	999.0000	1.0593	-1260.2532	-1281.7533	43.0001	0.0000
NM_018368	XM_518573	LMBRD1	LMBR1 domain containing 1	1.4934	1.0260	-1125.8698	-1145.9865	40.2335	0.0000
NM_002858	XM_513575	ABCD3	ATP-binding cassette, sub-family D (ALD), member 3	999.0000	1.3483	-1628.5040	-1645.9934	34.9789	0.0000
NM_173611	XM_510292	FAM98B	family with sequence similarity 98, member B	1.0182	3.1674	-1837.4997	-1852.6374	30.2754	0.0000
NM_004751	XM_510451	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type	1.9523	1.0454	-2536.4624	-2551.2513	29.5779	0.0000
NM_152313	XM_522147	SLC36A4	solute carrier family 36 (proton/amino acid symporter), member 4	999.0000	1.1599	-1323.4990	-1337.0865	27.1751	0.0000
NM_032604	XM_525719	ABHD1	abhydrolase domain containing 1	999.0000	4.7125	-1336.6032	-1347.3059	21.4054	0.0002
NM_016072	XM_520789	GOLT1B	golgi transport 1 homolog B (S. cerevisiae)	676.2431	999.0000	-196.4478	-206.8443	20.7930	0.0002
NM_017837	XM_513236	PIGV	phosphatidylinositol glycan, class V	999.0000	1.6995	-627.0900	-637.1605	20.1409	0.0003
NM_001001922	XM_521793	OR52N5	olfactory receptor, family 52, subfamily N, member 5	1.1951	1.8881	-635.9808	-643.9080	15.8544	0.0005
NM_173514	XM_517758	FLJ90709	hypothetical protein FLJ90709	999.0000	1.2236	-873.2634	-882.2775	18.0281	0.0008
NM_005819	XM_514037	STX6	syntaxin 6	2.3478	1.2598	-1142.3571	-1150.1565	15.5988	0.0023
NM_001774	XM_512814	CD37	CD37 antigen	1.5244	999.0000	-981.1560	-987.3475	12.3832	0.0025

Table 2.1 (Continued) | Genes with $d_N/d_S > 1$ on both the human and the chimpanzee lineages ($LRT_{free\text{-}ratio}$, $P < 0.05$). Genes that are also found following $d_S = 0$ filter step is shaded in light grey.

Human RefSeq	Chimp RefSeq	HUGO Symbol	Gene Description	Human d_N/d_S	Chimp d_N/d_S	$LRT_{free\text{-}ratio}$	$LRT_{one\text{-}ratio}$	$2\Delta\ell$	P value
NM_178498	XM_508335	SLC5A12	solute carrier family 5 (sodium/glucose cotransporter), member 12	999.0000	1.1010	-559.0200	-566.3457	14.6514	0.0035
NM_004393	XM_516461	DAG1	dystroglycan 1 (dystrophin-associated glycoprotein 1)	999.0000	24.6796	-591.6718	-598.6061	13.8686	0.0050
NM_000341	XM_515443	SLC3A1	solute carrier family 3, member 1	1.2086	427.4631	-3390.3119	-3395.4993	10.3746	0.0064
NM_031284	XM_510657	ADPGK	ADP-dependent glucokinase	999.0000	1.0063	-617.2442	-623.5232	12.5580	0.0091
NM_017801	XM_521192	CMTM6	CKLF-like MARVEL transmembrane domain containing 6	999.0000	1.8925	-678.8722	-684.8626	11.9807	0.0118
NM_002641	XM_520945	PIGA	phosphatidylinositol glycan, class A	999.0000	999.0000	-950.7881	-956.6932	11.8101	0.0127
NM_001877	XM_514158	CR2	complement component (3d/Epstein Barr virus) receptor 2	206.7022	1.0219	-5072.3451	-5078.1223	11.5543	0.0142
NM_020683	XM_513638	ADORA3	adenosine A3 receptor	999.0000	1.1132	-937.1228	-942.8838	11.5220	0.0144
NM_001005184	XM_524918	OR6K6	olfactory receptor, family 6, subfamily K, member 6	999.0000	999.0000	-267.9036	-273.6509	11.4946	0.0145
NM_001004713	XM_524139	OR1I1	olfactory receptor, family 1, subfamily I, member 1	999.0000	999.0000	-622.6543	-628.3556	11.4027	0.0151
NM_001029998	XM_526698	C4orf13	chromosome 4 open reading frame 13	920.9490	999.0000	-134.0308	-139.5582	11.0548	0.0176
NM_004854	XM_515653	CHST10	carbohydrate sulfotransferase 10	999.0000	958.3803	-1349.8424	-1355.3277	10.9708	0.0182
NM_000870	XM_518024	HTR4	5-hydroxytryptamine (serotonin) receptor 4	999.0000	999.0000	-421.9822	-426.9837	10.0030	0.0281
NM_181093	XM_513987	SCYL3	SCY1-like 3 (S. cerevisiae)	999.0000	999.0000	-2042.7600	-2047.7024	9.8847	0.0296
NM_003950	XM_512484	F2RL3	coagulation factor II (thrombin) receptor-like 3	999.0000	1.0415	-747.8940	-752.7965	9.8049	0.0308
NM_032871	XM_522102	TNFRSF19L	tumor necrosis factor receptor superfamily, member 19-like	3.0132	1.5004	-1129.7608	-1134.6585	9.7953	0.0309
NM_001005189	XM_524912	OR6Y1	olfactory receptor, family 6, subfamily Y, member 1	1.8015	1.5078	-639.2282	-642.3898	6.3233	0.0466

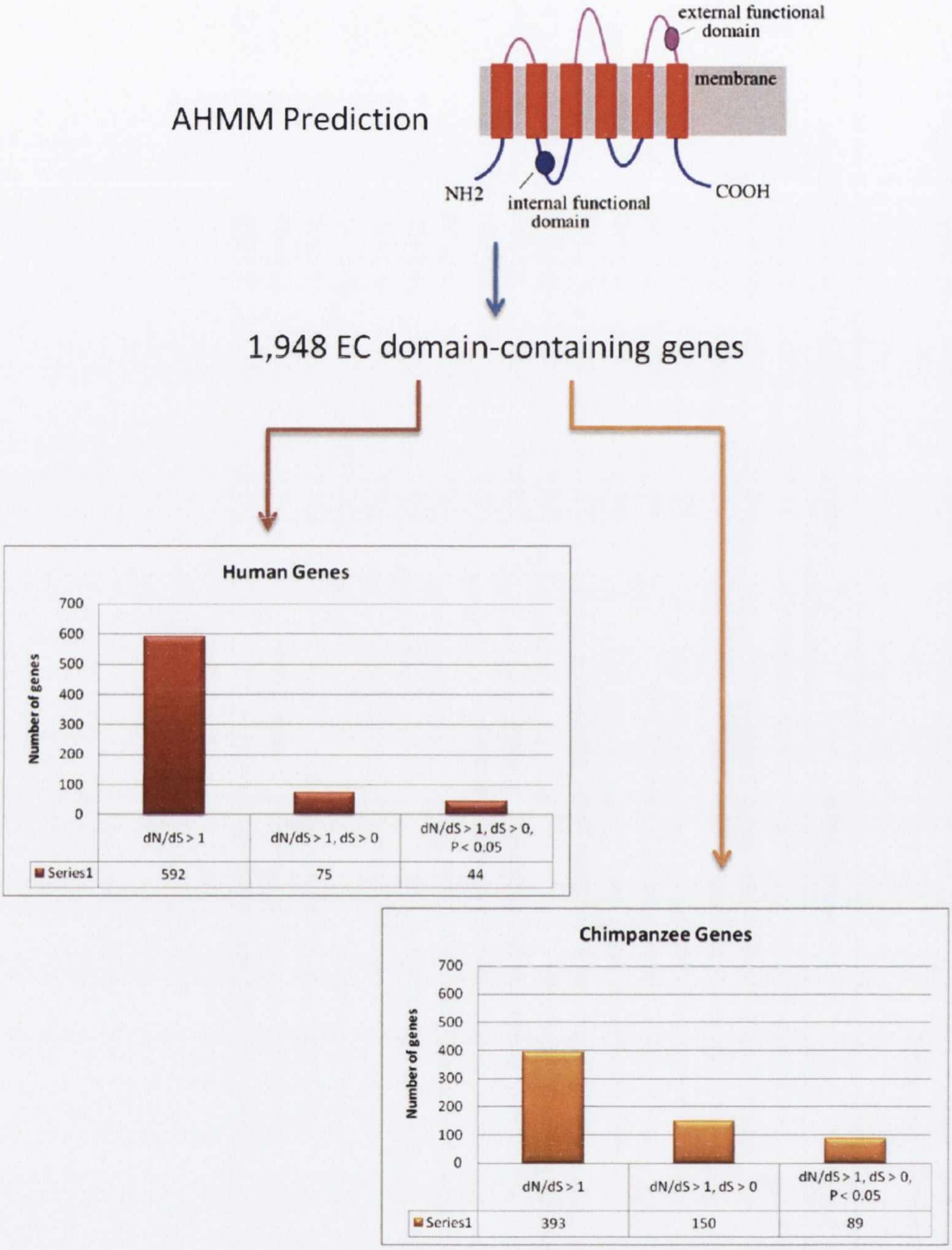


Figure 2.2 | Summary of AHMM implementation on HCM orthologous trios. A total of 592 human and 393 chimpanzee genes were found to have $d_N/d_S > 1$. Following filtering for $d_S = 0$, the numbers were reduced to 75 and 150 respectively. Of these, 44 human and 89 chimpanzee genes were found to be significant under the LRT tests conducted.

Each of these 44 genes was also tested under another model in which the d_N/d_S ratio was allowed to vary only on the human lineage (NSsites = 0, model = 2). A LRT comparing this more stringent test of positive selection on the human lineage to the one-ratio model revealed 22 genes as having a signature of positive selection.

Given the draft nature of the chimpanzee genome and the gene predictions based on it, it was of concern that poor quality alignments and/or misprediction of exons could lead to spurious signals of positive selection in some genes. To account for this, each alignment was manually checked and 7 of the 22 genes were excluded due to poor alignment quality and one gene had since been removed by NCBI (XM_497249). Undoubtedly, this problem is prevalent in other genomic analysis of positive selection using the chimpanzee genome. SACM1L was also excluded as it does not have an extracellular domain but rather is an endoplasmic reticulum membrane protein (Rohde et al. 2003). The 13 other genes with robust evidence for positive selection in their extracellular domains are implicated in processes including immunity, olfaction, and reproduction (Table 2.2).

Table 2.2 | Genes with $d_N/d_S > 1$ and $d_S > 0$ on the human lineage ($LRT_{model=2}$, $P < 0.05$).

Human RefSeq	HUGO Symbol	Gene Description	d_N/d_S	$LRT_{model=2}$	$LRT_{one\text{-}ratio}$	$2\Delta\ell$	P value
NM_173611	FAM98B	family with sequence similarity 98, member B	1.1831	-1837.2850	-1852.6374	30.7048	0.0000
NM_015879	ST8SIA3	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 3	20.8445	-1760.7695	-1770.2806	19.0222	0.0003
NM_021229	NTN4	netrin 4	1.0780	-3365.9847	-3371.5963	11.2231	0.0089
NM_001774	CD37	CD37 antigen	999.0000	-982.2955	-987.3475	10.1041	0.0093
XM_059074	LRRC38	leucine rich repeat containing 38	999.0000	-122.2891	-127.0493	9.5204	0.0112
NM_015187	KIAA0746	KIAA0746 protein	7.5097	-3999.7602	-4004.2016	8.8828	0.0115
NM_003259	ICAM5	intercellular adhesion molecule 5, telencephalin	999.0000	-3860.9715	-3865.4672	8.9915	0.0119
XM_040592	ZNF469	zinc finger protein 469	66.4466	-13831.4644	-13835.3435	7.7581	0.0157
NM_000341	SLC3A1	solute carrier family 3, member 1	544.1354	-3017.2702	-3020.9999	7.4594	0.0163
NM_183061	SLC9A10	solute carrier family 9, member 10	999.0000	-3351.9670	-3355.8508	7.7676	0.0167
NM_007267	TMC6	transmembrane channel-like 6	999.0000	-1949.3695	-1952.5548	6.3706	0.0269
NM_001005189	OR6Y1	olfactory receptor, family 6, subfamily Y, member 1	999.0000	-639.3123	-642.3898	6.1550	0.0288
NM_152481	FLJ25660	hypothetical protein FLJ25660	999.0000	-2418.9476	-2421.8955	5.8957	0.0318

2.3.2 Comparison of HCM Orthologues to Previous Studies

Using the HCM method, 13 extracellular domains of human genes based on AHMM prediction were found to be putatively under positive selection. Of these, two were previously analysed by Clark *et al.* but only one gene (NTN4) had a signature of positive selection in their analysis. Nielsen *et al.* examined 8 of the 13 genes, but none of these was predicted to be positively selected in their study. Comparison to the results from the CSAC showed that 7 of the 13 genes had been analysed in this dataset, and only one (Cluster of Differentiation 37, CD37) was found to have $K_A/K_I > 1$.

Although CD37 was identified by CSAC to have a signature of positive selection over the entire gene, following manual analyses of the genes putatively under positive selection, we have found that the effect is limited to the extracellular portion. CD37 has 12 nonsynonymous changes compared to chimpanzee and all of these are located in the extracellular domain. This over-representation of nonsynonymous changes in the extracellular region is statistically significant given the size of the domain and the size of the whole protein ($P < 0.05$). There are only 2 synonymous changes in this region, with a further 5 synonymous substitutions in the rest of the gene.

2.3.3 Pairwise Human-Chimpanzee d_N/d_S

The use of an outgroup species, although very useful in assigning the direction of selection, can result in many genes being excluded from the analysis due to difficulties in orthology assignment. In order to identify additional candidate genes that may have been missed in the previous analysis, pairwise human-chimpanzee d_N/d_S ratios were calculated under two models and compared by LRT.

Of 6,096 human-chimpanzee extracellular domain pairwise alignments, 1,120 had $d_N/d_S > 1$. Compared to the 733 genes with $d_N/d_S > 1$ that Nielsen *et al.* (Nielsen et al. 2005) had identified, we have found approximately 65% more genes with evidence of positive selection. Again, it should be noted that many of the genes with elevated d_N/d_S have very low d_S (Supplementary Table 2.2).

One of the drawbacks of examining particular domains compared to the entire gene sequence is that the power of the LRT to detect positive selection is significantly reduced, most likely due to both the short alignment length and the few lineages examined (Anisimova, Bielawski, and Yang 2001). Indeed, in only 20 genes with $d_N/d_S > 1$ is the signature of positive selection supported by the LRT ($P < 0.05$). Two genes (XM_374653, XM_497314) have since been removed as predicted genes as part of the ongoing annotation of the human genome by NCBI. We have also excluded a further two genes (SLC16A4, GOLT1B) due to alignment quality issues (Table 2.3).

In many cases low d_S results in estimates of d_N/d_S that are extremely elevated or equal to infinity. It is perhaps more informative in these cases to examine the actual number of nonsynonymous and synonymous substitutions. In the 16 remaining genes with relatively robust evidence for positive selection, there is an average of seven nonsynonymous substitutions and virtually no synonymous substitutions (CD34 has one) in the extracellular domains, strongly supporting the case for positive selection in these genes. In most incidences, the majority or all of the nonsynonymous changes within these genes are located in this region (Table 2.4).

Table 2.3 | Human genes with $d_N/d_S > 1$ ($P < 0.05$) in pairwise alignments with chimpanzee orthologues.

Human RefSeq	HUGO Symbol	Gene Description	Human d_N/d_S	$LRT_{\omega\ ML}$	$LRT_{\omega\ fixed}$	$2\Delta\ell$	P value
NM_181093	SCYL3	SCY1-like 3 (S. cerevisiae)	99.0000	-1374.4799	-1377.5160	6.0721	0.0088
XM_496025	VSIG7	V-set and immunoglobulin domain containing 7	6.7013	-1238.7734	-1241.6480	5.7493	0.0111
NM_001025109	CD34	CD34 antigen	99.0000	-1216.2808	-1218.7802	4.9989	0.0160
NM_006781	C6orf10	chromosome 6 open reading frame 10	99.0000	-2110.8394	-2113.2166	4.7544	0.0185
NM_054030	MRGPRX2	MAS-related GPR, member X2	99.0000	-359.2835	-361.6509	4.7348	0.0187
NM_003555	OR1G1	olfactory receptor, family 1, subfamily G, member 1	99.0000	-472.4164	-474.7449	4.6570	0.0193
NM_032604	ABHD1	abhydrolase domain containing 1	99.0000	-885.6155	-887.8744	4.5178	0.0203
NM_001001958	OR7G3	olfactory receptor, family 7, subfamily G, member 3	99.0000	-402.5083	-404.7239	4.4311	0.0207
NM_138337	CLEC12A	C-type lectin domain family 12, member A	99.0000	-820.7774	-822.9799	4.4050	0.0209
NM_032498	PEPP-2	PEPP subfamily gene 2	5.6323	-853.0307	-855.2109	4.3605	0.0213
NM_003778	B4GALT4	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 4	99.0000	-1285.7929	-1287.8923	4.1989	0.0224
NM_000873	ICAM2	intercellular adhesion molecule 2	99.0000	-920.1275	-921.8713	3.4875	0.0342
NM_001004724	OR4N5	olfactory receptor, family 4, subfamily N, member 5	99.0000	-386.5319	-388.2362	3.4086	0.0382
NM_001005338	OR5H1	olfactory receptor, family 5, subfamily H, member 1	99.0000	-429.1760	-430.7892	3.2264	0.0043
NM_002002	FCER2	Fc fragment of IgE, low affinity II, receptor for (CD23A)	99.0000	-784.3742	-785.9843	3.2202	0.0447
NM_198085	RNF148	ring finger protein 148	99.0000	-794.6052	-796.1973	3.1843	0.0461

Table 2.4 | Number of nonsynonymous and synonymous substitutions in extracellular domains and over the full gene length. (N_{EC} - non-synonymous substitutions in extracellular domain; S_{EC} - synonymous substitutions in extracellular domains; N_{full} - non-synonymous substitutions over full gene length; S_{full} - synonymous substitutions over the full gene length).

Human RefSeq	HUGO Symbol	Gene Description	N_{EC}	S_{EC}	N_{full}	S_{full}
NM_181093	SCYL3	SCY1-like 3 (S. cerevisiae)	8	0	10	4
XM_496025	VSIG7	V-set and immunoglobulin domain containing 7	9	0	9	0
NM_001025109	CD34	CD34 antigen	10	1	10	3
NM_006781	C6orf10	chromosome 6 open reading frame 10	9	0	9	0
NM_054030	MRGPRX2	MAS-related GPR, member X2	8	0	13	2
NM_003555	OR1G1	olfactory receptor, family 1, subfamily G, member 1	8	0	13	4
NM_032604	ABHD1	abhydrolase domain containing 1	7	0	9	0
NM_001001958	OR7G3	olfactory receptor, family 7, subfamily G, member 3	5	0	8	4
NM_138337	CLEC12A	C-type lectin domain family 12, member A	7	0	9	0
NM_032498	PEPP-2	PEPP subfamily gene 2	13	0	19	0
NM_003778	B4GALT4	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 4	6	0	7	1
NM_000873	ICAM2	intercellular adhesion molecule 2	6	0	6	0
NM_001004724	OR4N5	olfactory receptor, family 4, subfamily N, member 5	5	0	6	2
NM_001005338	OR5H1	olfactory receptor, family 5, subfamily H, member 1	4	0	7	8
NM_002002	FCER2	Fc fragment of IgE, low affinity II, receptor for (CD23A)	5	0	6	0
NM_198085	RNF148	ring finger protein 148	5	0	13	0

2.3.4 HCR Extracellular Domain-Containing Proteins

2.3.4.1 The RMGSAC – Analysis of HCR Orthologous Trios

In 2007, the Rhesus Macaque Genome Sequencing and Analysis Consortium (RMGSAC) published their findings following a series of analyses of the genome sequence of an Indian-origin *Macaca mulatta*, including an analysis to detect positive selection (Gibbs et al. 2007).

Their analysis pipeline produced 10,375 orthologous HCR trios and these were analysed using PAML to calculate d_N/d_S ratios. LRTs for several different models were performed including (a) a test for positive selection across all branches of the phylogeny (TA test), (b) a test for positive selection on the human branch (TH test), (c) a test for positive selection on the chimpanzee branch (TC test), and (d) a test for positive selection on the macaque branch (TM test). They identified, respectively for the afore-mentioned tests, 67, 2, 14 and 131 genes to be under positive selection. The sets overlapped by 36 genes, giving a total of 178 genes putatively under positive selection.

The genes under selection were found to be enriched in previously reported categories including defense and immunity, signal transduction and cell adhesion. Additionally they found enrichment of several new categories including ion binding and oxireductase activity. On the other hand, they found only weak enrichment for genes implicated in apoptosis and spermatogenesis. These categories of genes have previously been found to be under strong positive selection.

2.3.4.2 Bakewell et al. – HCR Analysis via Branch-Site Likelihood Method

At the same time, Bakewell *et al.* reported their findings, that there were more chimpanzee genes that underwent positive selection than human. Using a conservative dataset of 13,888 HCR trios, they applied branch-site likelihood method and identified 154 and 233 genes under positive selection in human and chimpanzee lineages respectively. Following Bonferroni correction to control for multiple testing, the numbers were reduced to 2 and 21.

2.3.4.3 HCR Extracellular Domain Analysis

With the availability of the macaque genome, a new analysis utilising the extracellular domain method was conducted. A dataset of 4,745 human-chimpanzee-macaque orthologues with an AHMM-predicted extracellular domain was analysed. We found 1,096 human and 838 chimpanzee genes with $d_N/d_S > 1$.

Under the LRT comparison between the free-ratios and the one-ratio models, the number of genes reduced dramatically to 91 and 104 for human and chimpanzee, respectively. However, following the Benjamini and Hochberg false discovery rate correction for multiple testing (Benjamini and Hochberg 1995a), 2 human and 4 chimpanzee genes were detected at significance $P < 0.05$ (Supplementary Table 2.3). The human genes found were chloride channel 4 (CLCN4) and hyperpolarisation activated cyclic nucleotide-gated potassium channel 1 (HCN1); the former which physiological role remains unknown but speculated to contribute to the pathogenesis of neuronal disorders, the latter linked to rhythmic activity in both the heart and the brain.

2.4 Discussion

The detection of positive selection can be conducted by fitting models of evolution by maximum likelihood to multiple sequence alignments of orthologous coding sequences. Several studies have been carried out recently to detect positive selection in the human and chimpanzee genomes but they have tended to concentrate on cases where there is a significant signature of selection over the entire gene (Clark et al. 2003; CSAC 2005; Nielsen et al. 2005; Arbiza, Dopazo, and Dopazo 2006). These analyses are likely to have missed cases where positive selection does not act uniformly over a gene but rather is most prevalent in particular domains or regions of the gene.

We proposed the extracellular domain of proteins as likely targets of increased selective pressure. Focusing on this particular domain has the advantage of avoiding a signal of positive selection in this region being swamped by purifying selection elsewhere in the gene but at the same time, the method can also be limiting in that the shorter alignment lengths can result in reduced statistical power of the LRT. However, we have previously demonstrated that the power of detecting positive selection in the cell-surface protein, Cluster of Differentiation 2 (CD2), was increased when only the extracellular domain was analysed (Lynn et al. 2005). Using mouse as outgroup, 13 human genes with a robust signal of positive selection in their extracellular domain have been identified.

The pairwise alignments of human and chimpanzee extracellular domains generated additional genes that may be analysed but were previously excluded due to difficulties in orthology assignment in our HCM orthologues dataset. However, without outgroup, the lineage on which selection has acted cannot be identified. Moreover, due to the shorter alignment lengths of extracellular domains, the LRT frequently lacks power to detect positive selection. Nevertheless, this dataset of genes with $d_N/d_S > 1$ remains enriched for true cases of positive selection that could be confirmed in more detailed analyses. Despite the drawback of low power of the LRT, a further 16 genes were identified to have robust signatures of positive selection, of which 10 had not been identified in previous genomic comparisons of humans and chimpanzees.

The 29 genes identified using either the mouse outgroup or pairwise d_N/d_S approach as having a significant signature of positive selection in their extracellular encoding regions include genes involved in processes similar to those previously found (Clark et al. 2003; Nielsen et al. 2005).

A number of olfactory genes (OR1G1, OR4N5, OR5H1, OR6Y1, OR7G3) were detected, and this is a category of genes that had frequently been found to be subject to positive selection (Gilad et al. 2003; Gilad, Man, and Glusman 2005). It should be noted that some of the olfactory receptors under positive selection may have been recently pseudogenised, although there is evidence that at least 3 of them (OR1G1, OR5H1, OR7G3) are at least still expressed in humans.

Three genes (NTN4, ST8SIA3, ICAM5) have been found to play roles in neuronal and brain development (Angata et al. 2000; Koch et al. 2000). MRGPRX2, another gene which emerged from the pairwise analysis, is involved in pain sensation. 8 of the 13 nonsynonymous changes between human and chimpanzee in this gene are located in the extracellular domain, significantly more than expected given the length of this domain (82 amino acids) ($P < 0.05$). Yang *et al.* have also recently demonstrated positive selection in this gene (Yang et al. 2005).

Genes expressed in the testis have also been reported to have a strong tendency to be subject to positive selection in a number of studies (Khaitovich et al. 2005; Nielsen et al. 2005). Three of the genes identified in this study are either selectively expressed (SLC9A10, PEPP-2, C6Orf10) or highly expressed (ABHD1) in the testis (Liang et al. 1994; Wayne et al. 2002; Edgar 2003). PEPP-2 was also found to be subject to positive selection by Nielsen *et al.* (Nielsen et al. 2005), although we have found the majority of nonsynonymous substitutions in the PEPP-2 extracellular domain (Table 2.4).

A number of other genes identified from the analyses are cell-surface receptors (CD34, CD37, ICAM2, FCER2(CD23), CLEC12A, VSIG7) with roles in immunity and/or cell adhesion. CD34, CD37 and ICAM2 have previously been identified as having undergone adaptive evolution on the human lineage (Clark et al. 2003; CSAC 2005). CD37 antigen is a leukocyte-specific tetraspanin protein (Horejsi and Vlcek 1991) that is implicated in the regulation of T-cell proliferation as well as B-cell responses (van Spriel et al. 2004). While the function of CD37 remains unclear, its expression specificity would suggest specialised role within the immune system. All 12 nonsynonymous substitutions in CD37 occur in the extracellular domain. Similarly, for CD34 and ICAM2, all the nonsynonymous changes are found in the extracellular domain (or signal peptide), indicating that it is this region in particular that has been subject to adaptive evolution in these genes. The other receptors (FCER2, CLEC12A and VSIG7) also have the majority or all of their nonsynonymous substitutions in this region.

We speculate that the selective pressure driving this effect in these genes to be a result of a host-pathogen genetic conflict. Interestingly, CD81, a tetraspanin protein related to CD37, has been shown to bind the Hepatitis C virus in the same region (Bertaux and Dragic 2006). FCER2 (CD23) is the low affinity receptor for immunoglobulin E (IgE), which plays a critical role in parasite immunity and allergies (Kijimoto-Ochiai 2002). The engagement of FCER2 is implicated in the cytoadherence of the malaria parasite, *Plasmodium falciparum*, which significantly impacts the pathogenesis of the cerebral infection (Pino et al. 2004). It is widely held that malaria has been one of the most significant selective forces in human history and a number of other genes, most classically, the sickle-cell mutation in the Haemoglobin-B genes, have been shown to be subject to strong selection (Sabeti et al. 2006).

Further circumstantial evidence of host-pathogen interactions driving change in these cell surface receptors is provided by a recent study proposing a binding activity of SARS Coronavirus X4 protein to the integrin I (extracellular) domain of ICAM2 – a cell surface adhesion molecule – or other similar molecules (Hanel et al. 2006). Certainly, SARS as a recently emergent virus cannot be the selective force in question, but this raises the likelihood of similar interactions by other viruses with a longer history in human populations.

Another gene identified in this study with implications for susceptibility to infectious diseases is TMC6. Mutations in this gene are associated with a rare autosomal disorder, epidermodysplasia verruciformis, where patients have an abnormal susceptibility to human papilloviruses that are harmless to the majority of the population (Tate et al. 2004). Mutations in another of the genes, SLC3A1, an amino acid transporter, have been shown to cause another autosomal recessive disease called cystinuria, a condition affecting kidney function and resulting in increased renal infections (Calonge et al. 1995).

Several of the genes identified (SCYL3, RNF148, B4GALT4, FLJ25660, KIAA0746, ZFN469, FAM98B, LRRC38) have relatively little known of their biological functions. Perhaps studies such as this one will provide the impetus for detailed investigations of the biological roles of these genes as they have clearly been of importance in our evolution.

Combining the lists of genes of this chapter for Gene Ontology (GO) analysis, 126 biological process and 141 molecular function terms had significant MWU *P*-values ($P < 0.05$), but following Benjamini and Hochberg multiple correction, the numbers were reduced to 29 and 44 respectively. We found enrichment of the following GO molecular function terms among the genes detected to be under positive selection, including receptor activity, olfactory receptor activity, melanocortin receptor activity, immunoglobulin E binding, complement receptor activity, chemokine activity, interleukin-10 receptor activity and G-protein coupled receptor activity. In terms of GO biological processes, the categories over-represented include sensory perception of smell, response to stimulus, immune response, chemotaxis, inflammatory response, amino acid transport, natural killer cell activation and cellular defense response (Supplementary Table 2.4 - significant terms are shaded in blue, while significant terms following multiple correction are shaded in pink). This further supports our hypothesis that the extracellular domain of immunity-related genes may be subject to positive selection due to their interactions with pathogens and exposure to external environments.

The search for extracellular domains under positive selection using human-chimpanzee-macaque orthologous sequences, however, did not yield significant results on the human lineage. Published studies by the Rhesus Macaque Genome Sequencing and Analysis Consortium (RMGSAC) (Gibbs et al. 2007) and Bakewell *et al.* also failed to identify more than 2 genes on the human lineage, indicating it is not just an issue in detecting positive selection on the extracellular domain but in genes on the human lineage in general. It must also be note that the 2 genes identified by RMGSAC had a false discovery rate (FDR) of < 0.1 while the 2 genes identified by Bakewell *et al.* used FDR < 0.05. Under FDR < 0.1, depending on the multiple testing correction method applied, our HCR trio strategy did yield a small number of genes in which the extracellular domains were putatively under positive selection.

Table 2.5 | Summary of human and chimpanzee genes with $d_N/d_S > 1$ at $P < 0.1$ and $P < 0.05$.

	Method of Correction	**FDR**	**Number of Human Genes**	**Number of Chimpanzee Genes**
RMGSAC	Benjamini & Hochberg	0.10	2	14
Bakewell *et al.*	Bonferroni	0.05	2	21
This study	Benjamini & Hochberg	0.05	2	4
		0.10	3	5
This study	Bonferroni	0.05	0	2
		0.10	0	2

The two published studies, as well as our own, all showed more adaptive changes in chimpanzee genes than in human genes. Table 2.5 summarises the number of genes detected from all three studies, as well as the multiple correction method for comparative purposes.

The paper published by Kosiol *et al.* in 2008 investigated for evidence of positive selection using six mammalian genomes – human, chimpanzee, rhesus macaque, mouse, rat and dog. Their method is perhaps the most successful to date, as they detected 400 genes under positive selective, for the dataset across all branches. However, for the lineage branching to hominids, only 7 cases of positive selection were detected.

Prior to availability of the macaque genome, it was widely believed that the inclusion of a species closely related to human and chimpanzee would aid in the detection of cases of positive selection. However, it appears now that the power to detect positive selection for individual primate branches is primarily weak because of the low levels of interspecies divergence. The inclusion of other non-primate mammals appears to be helpful in allowing a distinction in inferring branches among the primates that are under adaptive evolution.

3. Adaptive Evolution in the Bovine Genome

3.1 Adaptive Evolution in the Bovine Genome

Identification of the genes which have undergone adaptive evolution is no longer simply of interest to evolutionary biologists, who wish to gain insight into the mechanisms of the evolutionary process. Rather, it is now accepted that identifying genomic regions which have undergone positive selection may reveal the genes that engender the key biological differences between a species of interest and other related species. As discussed in the previous chapter, comparative analyses of the available primate genomes have provided new insight into human evolution. The new availability of the bovine genome presented the opportunity to extend these studies to investigate the evolution of cattle, an important domestic species, on a genomic scale.

3.1.1 Comparative Analysis of Bovine Genes from the Bovine Genome Project

Up to now, genome-wide interspecies comparative analyses to detect positive selection have been focused mainly on human, and to a large extent, primate lineages such as those of chimpanzee and rhesus macaque. Other organism species regularly used included mouse, rat and dog (Clark et al. 2003; Nielsen et al. 2005; Bakewell, Shi, and Zhang 2007; Kosiol et al. 2008). Lynn *et al.* in 2005 published the first genomics approach to detection of positive selection in the bovine lineage, through analysis of 3,190 cow, human, mouse and pig orthologous gene sequences. A total of 211 genes shown significant acceleration on the bovine lineage but only 6 genes were found to have $d_N/d_S > 1$ (Lynn et al. 2005). Of these 6 genes, CD2 was investigated in detail to identify the specific sites subject to adaptive evolution. All 6 genes (CD2, ART4, TYROBP, IL2, IL5 and IL13) were also subsequently analysed for signatures of selection in cattle populations (Freeman et al. 2008).

The recent availability of a carefully curated and validated set of bovine genes through the effort of the Bovine Genome Sequencing and Analysis Consortium (BGSAC) enabled us to implement several approaches to detect adaptive evolution on the bovine lineage (Elsik et al. 2009)[5]. Genome-wide comparative analysis between cow and six other mammalian species - human, mouse, rat, dog, platypus and opossum - was carried out by us as part of the BGSAC to search for candidate genes potentially under positive selection in the bovine lineage. The relationship between these seven species is shown in Figure 3.1.

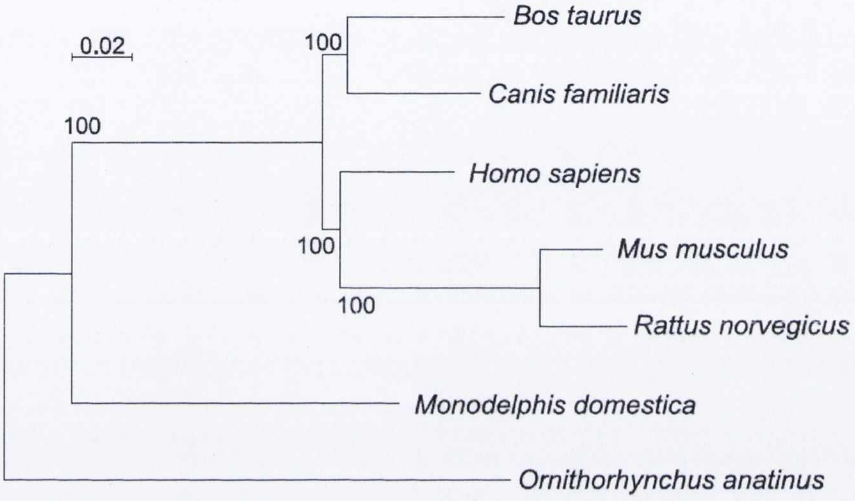


Figure 3.1 | Phylogenetic relationship between cow, dog, human, mouse, rat, platypus and opossum.

3.1.2 Extension of Extracellular Domain Analysis Protocol to Bovine Genes

The availability of earlier draft versions of the bovine genome, were also investigated in this thesis. Following the development of the protocol to detect positive selection in human and chimpanzee extracellular domains (based on human-chimpanzee-mouse orthologous trios), the method was extended to investigate for signatures of adaptive evolution in the extracellular domains of human, mouse, rat, dog and cow orthologous quintets.

[5] LPL Lau, DJ Lynn, DG Bradley are co-authors and participants of the Bovine Genome Consortium.

3.2 Materials and Methods

3.2.1 Human-Mouse-Rat-Dog-Cow (HMRDC) Orthology

Human, mouse, rat, dog and cow (HMRDC) gene entries were downloaded from the NCBI GenBank database (ftp://ftp.ncbi.nih.gov/genbank) (Benson et al. 2006). CODERET (Mullan and Bleasby 2002) was used to extract a total of 29,549 human, 50,309 mouse, 24,079 rat, 33,723 dog and 33,538 cow Reference Sequence (RefSeq) genes from the GenBank entries. To obtain the dataset of orthologous genes for all five species, the same BlastP and GeneID identification protocol described in Section 2.2.1 was applied. This resulted in five lists of best blast hits, one for each species, which were subsequently loaded into a MySQL database to query for 5-way reciprocal best hits.

3.2.2 HMRDC Extracellular Domain Analysis

The human AHMM extracellular domain dataset from the human-chimpanzee-mouse trio analysis were used as template sequences to extract extracellular domain regions in HMRDC quintets. The sequences were aligned as described in the previous chapter and CODEML from the PAML 3.14 suite of programs was used to calculate d_N/d_S ratios (Yang 1997).

To identify signatures of positive selection on each lineage, two models were tested - the one-ratio model and the free-ratio model - both of which have previously been described in Section 2.2.4. Additionally, a model specifying the bovine lineage as the “foreground” lineage (model2) was also applied where d_N/d_S was specifically allowed to vary unconstrained on this lineage only. Model2 was tested against the one-ratio model using the Likelihood Ratio Test (LRT), which was calculated from twice the difference between $\ln L_{model2}$ and $\ln L_{one\text{-}ratio}$ (i.e. $2\Delta\ell$). This value was then compared to a χ^2 distribution in order to obtain the *P* values.

3.2.3 Bovine Genome Project (BGP) Orthology Assignment

The Bovine Genome Sequencing and Analysis Consortium (BGSAC) provided a dataset which inferred the orthologous relationships between genes in the cow, human, mouse, rat, dog, platypus and opossum genomes. Bovine genes were predicted using GLEAN, a tool for creating consensus gene models by integrating evidence (Elsik et al. 2007) from various sources and algorithms, including NCBI, Ensembl, Fgenesh, Fgenesh++, Geneid, SGP2, aligned proteins and ESTs. A detailed discussion of the method and criteria used to define bovine gene models, known as the official gene set (OGS) by the BGSAC, are available in the BGP Supplementary Material (See Supplementary Documentation). The sequences for the other species were obtained from Ensembl v45.

Orthology was inferred using the Smith-Waterman algorithm to perform all-against-all protein sequence similarity searches (Elsik et al. 2009). The longest predicted transcript per locus was retained as the representative coding sequence. Orthologous groups were then formed by: (i) grouping recently duplicated sequences with >97% identity within genomes to be treated subsequently as single sequences; (ii) identifying reciprocal best hits between genomes, and; (iii) expanding the seed orthologous groups by inclusion of co-orthologous sequences that are more similar to the orthologous gene than to any other gene in any other genome.

3.2.4 Reconstruction of Phylogenetic Tree for BGP Orthologues

Most modern computational methods implemented to detect evidence of positive selection require the provision of a phylogenetic tree for each orthologous set of genes. To date, genome projects have tended to investigate evidence of positive selection either in the analysis of pairwise alignments or in datasets of strict 1:1 orthologues in a small number of comparison species, in which the orthologue was present in all species examined (CSAC 2005; Lindblad-Toh et al. 2005; Gibbs et al. 2007). These analyses could use a single simple species tree to represent the phylogenetic relationship of each orthologous gene set. In this project, we took advantage of the increasing number of mammalian genome sequences available and analysed orthologues from seven species for evidence of adaptive evolution. Increasing the number of orthologous sequences in each alignment is expected to increase the power to detect positive selection (Anisimova, Bielawski, and Yang 2001).

While the BGP dataset contained orthologous genes from seven different species, not all orthologues were found in each of the seven genomes examined. As a result, the orthologous gene sets contained a variable number of sequences. This necessitated a reconstruction of a phylogenetic tree for each dataset individually. Neighbor-joining (NJ) phylogenetic trees were reconstructed using coding sequence alignments for each of the orthologous gene datasets using the neighbour algorithm implemented in PHYLogeny Inference Package (PHYLIP) (Felsenstein 2005).

3.2.5 Comparative Analysis of BGP Dataset

The CODEML program from PAMLv4 (Yang 2007) was used to perform maximum likelihood estimation of d_N/d_S for each gene from coding sequence alignments of each of the 10,519 orthologous groups. Similar to the previous studies carried out on extracellular domain dataset, a number of models were tested on the BGP dataset.

The one-ratio model acts as the null model (NSsites = 0, model = 0), where each lineage was modelled to have the same d_N/d_S ratio. The ratio is constrained between 0 and 1, and thus does not allow for the presence of positive selection. The free-ratio model (NSsites = 0, model = 1) allows independent d_N/d_S estimates for each of the lineages tested. The lineage-specific model (NSsites = 0, model = 2) was used as model of bovine-specific evolution, where the bovine lineage was selected as the “foreground” lineage and d_N/d_S was specifically allowed to vary unconstrained on this lineage only. As described in Section 3.2.2, these models were compared by LRT, with *P* values obtained through comparison to a χ^2 distribution.

3.3 Results

3.3.1 Extracellular Domain Analyses of HMRDC Quintets

Utilising gene models from early drafts of the bovine genome, 4,438 orthologous datasets of human, mouse, rat, dog and cow genes were identified. Of these, 1,325 contained an AHMM-predicted extracellular domain. 72 of these were excluded as there was insufficient sequence information for the CODEML analysis. Therefore, a total of 1,253 genes were tested.

In comparing the free-ratios model to the null model, 24 genes were found to have $d_N/d_S > 1$ on the bovine lineage. A LRT test revealed that for 5 of these genes, the model of variable selective pressure on the bovine lineage was statistically favoured ($P < 0.05$) (Table 3.1). A more stringent test of positive selection on the bovine lineage was also carried out, by using lineage-specific model2 which allows d_N/d_S ratio to vary only on the bovine lineage (the "foreground" lineage). A total of 18 genes were found to have $d_N/d_S > 1$. The majority of these genes had been previously detected in the free-ratio model test, except for 3 genes - SLC2A5, CNIH4 and SMPD1. 7 of the 18 genes were also found have statistically significant evidence of variable selective pressure ($P < 0.05$) (Table 3.2).

For the tests conducted, three genes were found to be under positive selection and have significant *P*-value in both result sets - SLC26A8, ATP6V0E1 and RAMP1.

All genes tested under both analyses with $d_N/d_S > 1$ were also analysed for Gene Ontology over-representation, with the molecular function and biological process most associated with involved in transmembrane transport activities and catabolic processes (Supplementary Table 3.1).

Table 3.1 | List of bovine genes with evidence of positive selection in the extracellular domain under the free-ratios model. Genes in Italic font with grey shading are also found to be under positive selection under the bovine lineage-specific model.

Bovine RefSeq	Human RefSeq	Gene Symbol	Gene Name	dN/dS	LRT Value	M1 P value	
XM_608983	*NM_052961*	*SLC26A8*	*solute carrier family 26, member 8*	*1.1406*	*24.977992*	*0.0003*	***
XM_868269	*NM_003945*	*ATP6V0E1*	*ATPase, H+ transporting, lysosomal 9kDa, V0 subunit e1*	*999*	*16.402968*	*0.0117*	*
XM_596577	NM_032422	GPR123	G protein-coupled receptor 123	999	15.36041	0.0176	*
XM_869191	*NM_005855*	*RAMP1*	*receptor (calcitonin) activity modifying protein 1*	*271.3277*	*13.02945*	*0.0426*	*
NM_001037593	NM_000837	GRINA	glutamate receptor, ionotropic, N-methyl D-aspartate-associated protein 1	999	12.85358	0.0454	*
XM_870670	NM_032609	COX4I2	cytochrome c oxidase subunit IV isoform 2 (lung)	999	12.223854	0.0572	
XM_615127	NM_020437	ASPHD2	similar to aspartate beta hydroxylase domain-containing 2	999	11.980954	0.0624	
NM_001001439	NM_001861	COX4I1	cytochrome c oxidase subunit IV isoform 1	999	11.871738	0.0649	
NM_174109	NM_000529	MC2R	melanocortin 2 receptor (adrenocorticotropic hormone)	1.4208	11.646596	0.0703	
XM_604234	NM_005927	MFAP3	microfibrillar-associated protein 3	1.3655	11.48978	0.0744	
XM_613630	NM_006320	PGRM2	progesterone receptor membrane component 2	1.639	8.85723	0.1818	
XM_867821	NM_153611	CYBASC3	cytochrome b, ascorbate dependent 3	1.1039	7.546412	0.2733	
NM_001038065	NM_002510	GPNMB	glycoprotein (transmembrane) nmb	999	7.326402	0.2917	
NM_205815	NM_006691	XLKD1	extracellular link domain containing 1	1.1269	6.327324	0.3875	
XM_588038	NM_080723	VMP	vesicular membrane protein p24	27.549	6.286686	0.3919	
XM_589432	NM_152310	ELOVL3	elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 3	1.0228	5.966446	0.4270	
XM_593491	NM_021181	SLAMF7	SLAM family member 7	1.3207	5.727644	0.4544	
XM_613965	NM_024843	CYBRD1	cytochrome b reductase 1	999	5.712138	0.4562	
XM_603485	NM_199133	LOC134145	hypothetical protein LOC134145	1.7939	4.912186	0.5551	
XM_614589	NM_001038603	MARVELD2	MARVEL domain containing 2	999	4.631858	0.5918	
NM_001008666	NM_015865	SLC14A1	solute carrier family 14 (urea transporter), member 1	1.4323	3.993804	0.6775	
NM_001034364	NM_033504	TMEM54	transmembrane protein 54	999	3.7349	0.7125	
NM_205798	NM_022152	TMBIM1	transmembrane BAX inhibitor motif containing 1	1.2595	3.261208	0.7754	
XM_590197	NM_001033517	SERINC4	serine incorporator 4	410.7366	1.129394	0.9802	

Table 3.2 | List of bovine genes with evidence of positive selection in the extracellular domain under the lineage-specific model2. Genes in Italic font with grey shading are also found to be under positive selection under the free-ratios model. Genes denoted (#) are genes not found when tested using free-ratios model to have $d_N/d_S > 1$.

Bovine RefSeq	Human RefSeq	Gene Symbol	Gene Name	dN/dS	LRT Value	M2 P value
XM_608983	*NM_052961*	*SLC26A8*	*solute carrier family 26, member 8*	*1.1173*	*13.645364*	*0.0002* ***
XM_615127	NM_020437	ASPHD2	similar to aspartate beta hydroxylase domain-containing 2	999	9.29899	0.0023 **
XM_868269	*NM_003945*	*ATP6V0E1*	*ATPase, H+ transporting, lysosomal 9kDa, V0 subunit e1*	*999*	*8.908564*	*0.0028* **
XM_604234	NM_005927	MFAP3	microfibrillar-associated protein 3	1.4953	7.04579	0.0079 **
XM_869191	*NM_005855*	*RAMP1*	*receptor (calcitonin) activity modifying protein 1*	*5.2636*	*6.879564*	*0.0087* **
NM_205815	NM_006691	XLKD1	extracellular link domain containing 1	1.162	4.578902	0.0324 *
XM_583977	NM_003039	SLC2A5	solute carrier family 2 (facilitated glucose/fructose transporter), member 5	999	3.874518	0.0490 *(#)
NM_001008666	NM_015865	SLC14A1	solute carrier family 14 (urea transporter), member 1	1.3812	3.733268	0.0533
XM_613965	NM_024843	CYBRD1	cytochrome b reductase 1	999	3.693298	0.0546
XM_593491	NM_021181	SLAMF7	SLAM family member 7	1.2666	3.233876	0.0721
XM_613630	NM_006320	PGRM2	progesterone receptor membrane component 2	1.5994	2.87564	0.0899
XM_867821	NM_153611	CYBASC3	cytochrome b, ascorbate dependent 3	1.4743	2.471276	0.1159
XM_589432	NM_152310	ELOVL3	elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 3	1.0457	2.189822	0.1389
NM_001034338	NM_014184	CNIH4	cornichon homolog 4 (Drosophila)	999	2.051128	0.1521 (#)
NM_001037593	NM_000837	GRINA	glutamate receptor, ionotropic, N-methyl D-aspartate-associated protein 1	999	1.938432	0.1638
XM_581333	NM_000543	SMPD1	sphingomyelin phosphodiesterase 1, acid lysosomal	1.3289	1.055878	0.3042 (#)
XM_603485	NM_199133	LOC134145	hypothetical protein LOC134145	1.5038	0.574974	0.4483
XM_590197	NM_001033517	SERINC4	serine incorporator 4	999	0.415582	0.5191

3.3.2 Positive Selection of Bovine Genes in BGP Dataset

To identify genes on the bovine lineage that have evidence of adaptive evolution, 10,519 bovine genes were compared to their putative orthologues (where present) in the human, mouse, rat, dog, opossum and platypus genomes. Of the 10,519 orthologous groups, 1,531 orthologues were missing in the human genome, 940 in mouse, 1,352 in rat, 1,895 in dog, 2,563 in opossum and 4,999 in platypus.

A total of 2,210 genes were identified to have evidence of variable selective pressure on the bovine lineage as determined by a statistically significant LRT where model 2 was significantly favoured (Elsik et al. 2009). Among them, 71 bovine genes were identified with $d_N/d_S > 1$ under model 2 and of these, 40 were also significant using the LRT and have statistically significant evidence of adaptive evolution on the bovine lineage (Table 3.3).

The bovine specific model described above is conservative in that it assumes that there is variable selective pressure only on the specified bovine lineage. It may be an unrealistic assumption that orthologous genes from the other divergent mammalian species are subject to almost uniform selective pressure. To overcome this assumption, the null one-ratio model was compared to another model, the free-ratios model, which allows variable selective pressure on all the lineages. An additional 16 of the 71 genes with $d_N/d_S > 1$ on the bovine lineage were found where the free-ratios model was significantly favoured ($P < 0.05$) (Table 3.4). (For additional result tables, see Supplementary Tables 3.2 to 3.5.)

Table 3.3 | List of genes identified under the lineage-specific model to have been subject to positive selection on bovine lineage identified as part of the analyses for the bovine genome project.

Bovine Gene ID	Human Ensembl ID	Gene Symbol	Gene Name	dN/dS	LRT Value	P Value
GLEAN_15922		Ctbs	chitobiase, di-N-acetyl-	1.0407	33.0474	0.001***
GLEAN_00966	ENSG00000187398	LUZP2	leucine zipper protein 2	1.0691	11.8997	0.001***
GLEAN_10454	ENSG00000205081	CXorf30	chromosome X open reading frame 30	1.1682	19.7457	0.001***
GLEAN_02457				1.7248	14.8805	0.001***
GLEAN_25881	ENSG00000126860	EVI2A	ecotropic viral integration site 2A	4.2355	17.59	0.001***
GLEAN_11540	ENSG00000170231	FABP6	fatty acid binding protein 6, ileal (gastrotropin)	1.0691	14.0517	0.001***
GLEAN_07078		Enc1	ectodermal-neural cortex 1	37.9468	62.2454	0.001***
GLEAN_25382	ENSG00000164304	CAGE1	cancer antigen 1	1.0635	16.515	0.001***
GLEAN_18851	ENSG00000176714	CCDC121	coiled-coil domain containing 121	1.0406	7.5234	0.01**
GLEAN_12455	ENSG00000006757	PNPLA4	patatin-like phospholipase domain containing 4	8.1573	6.6738	0.01**
GLEAN_07910		Ifnar2	interferon (alpha and beta) receptor 2	1.506	7.1265	0.01**
GLEAN_01453	ENSG00000169228	RAB24	RAB24, member RAS oncogene family	999	7.8645	0.01**
GLEAN_00418		Pla2g10	phospholipase A2, group X	999	7.3482	0.01**
GLEAN_07098	ENSG00000164136	IL15	interleukin 15	1.7933	8.65	0.01**
GLEAN_25210	ENSG00000156384	C10orf78	chromosome 10 open reading frame 78	1.049	7.769	0.01**
GLEAN_18597	ENSG00000129235	TXNDC17	thioredoxin domain containing 17	2.8219	8.8405	0.01**
GLEAN_22866		2310040G07Rik	RIKEN cDNA 2310040G07 gene	1.4752	8.0139	0.01**
GLEAN_25910	ENSG00000162493	PDPN	podoplanin	1.4412	7.8711	0.01**
GLEAN_20989	ENSG00000154035	IDBG-35547	transcript expressed during hematopoiesis 2	999	8.2142	0.01**
GLEAN_21922	ENSG00000178826	TMEM139	transmembrane protein 139	1.0638	7.6297	0.01**

Table 3.3 (Continued) | List of genes identified under the lineage-specific model to have been subject to positive selection on bovine lineage identified as part of the analyses for the bovine genome project.

Bovine Gene ID	Human Ensembl ID	Gene Symbol	Gene Name	dN/dS	LRT Value	P Value
GLEAN_07723	ENSG00000167088	SNRPD1	small nuclear ribonucleoprotein D1 polypeptide 16kDa	999	9.2869	0.01**
GLEAN_16592	ENSG00000107719	KIAA1274	KIAA1274	1.5404	7.6103	0.01**
GLEAN_13996	ENSG00000104408	EIF3E	eukaryotic translation initiation factor 3, subunit E	999	6.9977	0.01**
GLEAN_22648	ENSG00000183208	IDBG-29982	CDNA FLJ27017 fis, clone SLV05746.	999	4.227	0.05*
GLEAN_00853	ENSG00000179965	ZNF771	zinc finger protein 771	999	4.6139	0.05*
GLEAN_09957	ENSG00000168404	MLKL	mixed lineage kinase domain-like	1.298	5.0861	0.05*
GLEAN_05536		Lyl1	lymphoblastomic leukemia	1.6858	5.8985	0.05*
GLEAN_20841	ENSG00000174059	CD34	CD34 molecule	1.036	6.2553	0.05*
GLEAN_20249	ENSG00000100307	CBX7	chromobox homolog 7	999	4.1134	0.05*
GLEAN_02453	ENSG00000116127	ALMS1	Alstrom syndrome 1	999	5.0493	0.05*
GLEAN_05820	ENSG00000173401	GLIPR1L1	GLI pathogenesis-related 1 like 1	1.6137	6.5556	0.05*
GLEAN_12673	ENSG00000187942	LDLRAD2	low density lipoprotein receptor class A domain containing	1.4907	4.1599	0.05*
GLEAN_19856	ENSG00000161911	TREML1	triggering receptor expressed on myeloid cells-like 1	1.1303	5.2973	0.05*
GLEAN_19853	ENSG00000124731	TREM1	triggering receptor expressed on myeloid cells 1	1.1855	4.1069	0.05*
GLEAN_26048	ENSG00000161929	C17orf87	chromosome 17 open reading frame 87	1.5041	5.6321	0.05*
GLEAN_10642	ENSG00000125743	SNRPD2	small nuclear ribonucleoprotein D2 polypeptide 16.5kDa	3.1522	4.7035	0.05*
GLEAN_10081	ENSG00000132329	RAMP1	receptor (G protein-coupled) activity modifying protein 1	999	4.0506	0.05*
GLEAN_04862	ENSG00000137463	IDBG-38342	ovary-specific acidic protein	1.7336	5.1077	0.05*
GLEAN_19967	ENSG00000187837	HIST1H1C	histone cluster 1, H1c	999	4.3232	0.05*
GLEAN_15535	ENSG00000079435	LIPE	lipase, hormone-sensitive	999	5.7109	0.05*

Table 3.4 | List of genes identified under the free-ratios model to have been subject to positive selection on bovine lineage identified as part of the analyses for the bovine genome project.

Bovine Gene ID	Human Ensembl ID	Gene Symbol	Gene Name	dN/dS	LRT Value	M1 P Value
GLEAN_19128	ENSG00000188906	LRRK2	leucine-rich repeat kinase 2	999	0.4587	0.001***
GLEAN_23729		Il23r	interleukin 23 receptor	3.0384	0.7612	0.001***
GLEAN_06498	ENSG00000177875	IDBG-29906	CDNA FLJ32221 fis, clone PLACE6004005.	2.7955	2.6429	0.001***
GLEAN_14741	ENSG00000181355	OFCC1	orofacial cleft 1 candidate 1	1.3993	0.6134	0.001***
GLEAN_00283	ENSG00000179639	FCER1A	Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide	1.0067	1.6842	0.001***
GLEAN_08455	ENSG00000141380	SS18	synovial sarcoma translocation, chromosome 18	999	2.7775	0.01**
GLEAN_11591	ENSG00000160345	C9orf116	chromosome 9 open reading frame 116	999	2.6297	0.01**
GLEAN_08191	ENSG00000185942	FAM77D	Na+/K+ transporting ATPase interacting 3	999	1.9647	0.01**
GLEAN_03300	ENSG00000197858	GPAA1	glycosylphosphatidylinositol anchor attachment protein 1 homolog	999	1.4339	0.01**
GLEAN_20621		Leap2	liver-expressed antimicrobial peptide 2	999	0.7322	0.01**
GLEAN_15561		Ubxd6	UBX domain containing 6	1.4022	2.0273	0.01**
GLEAN_25653	ENSG00000162892	IL24	interleukin 24	1.1881	2.563	0.01**
GLEAN_23649		Zfp580	zinc finger protein 580	999	2.2001	0.05*
GLEAN_24551	ENSG00000178385	IDBG-79678	pleckstrin homology domain containing, family M, member 1-like	999	1.1131	0.05*
GLEAN_09242	ENSG00000127362	TAS2R3	taste receptor, type 2, member 3	1.1784	3.7418	0.05*
GLEAN_25096		1600029D21Rik	RIKEN cDNA 1600029D21 gene	1.1734	1.6537	0.05*

3.4 Discussion

We previously proposed that the extracellular domains of proteins are likely targets of adaptive evolution, and tested this hypothesis on various primate lineages. With the human, mouse, rat, dog and cow orthologous dataset, the same protocol was applied but with an aim to identify such signatures on the bovine lineage.

Under the free-ratios model, where every lineage was allowed to have varied rate of evolution, 24 genes were found to have $d_N/d_S > 1$ on the bovine lineage (Table 3.1). Among them, five were also statistically significant in comparison to the null model of no variable selective pressure. When a more stringent lineage-specific model2 was applied, 18 genes were identified with $d_N/d_S > 1$ (Table 3.2). Fifteen of these had already been detected in the free-ratios analysis. Seven of these 18 genes were statistically significant and three of them corresponded to three of five genes found in the free-ratios analysis. The level of concordance between the two result sets indicated a relatively solid approach in the detection of positive selection based on the extracellular domain. In total, nine genes were found to be under positive selection on the bovine lineage.

Two solute carriers were among those detected. SLC26A8, a sulphate/anion transporter, is primarily expressed in the spermatocytes, and while a mutation of SLC26A8 was not shown to cause infertility in human (Makela et al. 2005), it has recently been implicated in affecting sperm motility (Lhuillier et al. 2009). SLC2A5, on the other hand, is a facilitated glucose/fructose transporter which has been found to be present in high level in human testis and spermatozoa. Fructose intake is thought to prevent premature activation of the spermatocyte (Burant et al. 1992). SLC2A5 is also involved in fructose delivery at the lumen of small intestine and a deficiency of this carrier causes fructose malabsorption (Barone et al. 2009).

ATP6V0E1 is an essential component of vacuolar ATPase (V-ATPase) proton pump (Blake-Palmer et al. 2007), which mediates acidification of eukaryotic intracellular organelles. Another gene that has a function related to cellular transport is RAMP1. RAMP1 is required for the transportation of calcitonin-receptor-like receptor (CRLR) to the plasma membrane (McLatchie et al. 1998).

XLKD1, alternatively known as lymphatic vessel endothelial hyaluronan receptor 1 (LYVE1), encodes a glycoprotein which acts as a receptor to hyaluronan, a chief component of the extracellular matrix. Hyaluronan, or hyaluronic acid, is also a component of the extracellular capsule of group A streptococcus and has been reported to be a virulence factor (Wessels et al. 1991; Moses et al. 1997). GRINA (or TMBIM3) is a member of Bax-inhibitor 1 family which previous genomic comparative study showed that it may participate in cell death pathways through promotion of tumour metastasis (Zhou et al. 2008). The function of microfibrillar-associated protein 3 (MFAP3) and aspartate beta-hydroxylase domain-containing protein 2 (ASPHD2) are, as yet, not fully explored.

There are a number of other genes from the extracellular domain analyses that are potentially interesting for follow-up studies, such as SLAMF7 which is implicated in an immunomodulatory role and MC2R which is a receptor specific for adrenocorticotropic hormone (ACTH).

Gene Ontology analysis of extracellular-domain contaning genes found with $d_N/d_S > 1$ reveals that the biological processes this set are over-represented are that of transport processes (including fructose transport and oxalate transport) and catabolic processes (including glycosaminoglycan catabolic process and sphingomyelin catabolic process). In terms of molecular function, transmembrane transporter activities are observed for oxalate, sulphate, urea and fructose, along with adrenocorticotropin receptor activity, coreceptor soluble ligand activity, sphingomyelin phosphodiesterase activity and ligase activity. These genes are most associated for localisation in the membrane, mitochondrial respiratory chain complex and tight junction. (Supplementary Table 3.1)

In investigating genes under adaptive evolution from the dataset obtained for the BGP, more distant outgroup species were added, namely the platypus and the opossum. However, orthologous genes could not be identified in all cases. Consequently, each gene was analysed with an independently reconstructed phylogenetic tree based on the sequences available, which may or may not conform to the accepted general phylogeny of mammals.

Genes that have $d_N/d_S > 1$ on the bovine lineage and are significant under either model 2 or the free-ratios model were found to be associated with a range of Gene Ontology biological processes including the immune response (IL24, IL15, IL23R, LEAP2, TREM1), cell adhesion (CD34), transcription (CBX7, HIST1H1C, ZNF771), and lipid metabolism (FABP6, LIPE, PNPLA4).

To determine if any particular functional categories were significantly associated with genes subject to positive selection we have mapped via orthology, bovine genes to human molecular function and biological process terms from the Gene Ontology (Ashburner et al. 2000) and PANTHER databases (Mi et al. 2007). For each ontology term, the distribution of log likelihood ratios associated with genes mapped to the term was compared, using a one-sided Mann-Whitney U (MWU) test, to the distribution of all log likelihood ratios, similarly to the method previously described (Gibbs et al. 2007). This approach has the potential to identify categories of genes that have a tendency towards being subject to positive selection despite the majority of genes not having stringent evidence of positive selection.

108 Gene Ontology molecular function and 130 Gene Ontology biological process terms had a significant MWU *P* values ($P < 0.05$), however, only five molecular function terms were significant (FDR < 0.1) after correction for multiple testing using the Benjamini and Hochberg correction for the false discovery rate (FDR) (Benjamini and Hochberg 1995b) (Supplementary Tables 3.6 and 3.7). Only two PANTHER ontology terms (Glutamate receptor and Cation transport) are significant after correction (Supplementary Table 3.8).

To investigate what pathways were represented in genes that had evidence of positive selection, bovine genes were mapped to pathway annotation via their human orthologues using InnateDB (Supplementary Table 3.9). InnateDB is a platform to facilitate systems level analysis, which integrates pathway and molecular interaction data from the major publicly available databases (Lynn et al. 2008b). No overrepresentation of any particular pathway was apparent in this dataset.

InnateDB was also used to investigate and visualize the molecular interaction networks of the genes which had evidence of positive selection and their interacting partners (Figure 3.2). Two different pairs of genes, with $d_N/d_S > 1$ on the bovine lineage, were observed to interact with each other. SNRPD1 and SNRPD2, which are both small nuclear ribonucleoproteins and are involved in spliceosome assembly, were found to be interacting partners. An interaction between glycosylphosphatidylinositol anchor attachment protein 1 (GPAA1) and the eukaryotic translation initiation factor (EIF3E) was also observed but it should be noted that this interaction is only supported by a single yeast 2-hybrid experiment.

Overall, in our effort to detect for positive selection, with emphasis on the bovine lineage, a number of categories of genes have been identified including immune-related functions, spermatogenesis, developmental, transcription and sensory perception.

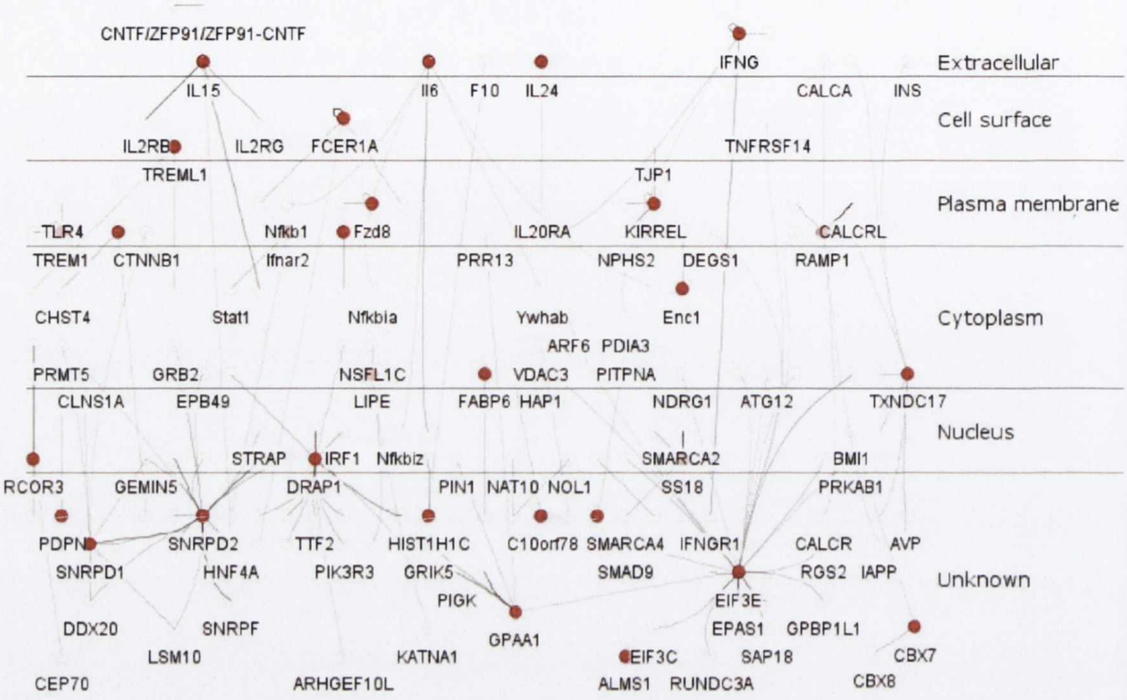


Figure 3.2 | Visualisation of the BGP genes under positive selection and their interactions in a molecular interaction network. Genes with $d_N/d_S > 1$ on the bovine lineage are represented by the red dots and their interacting partners are as identified in the InnateDB database.

As both the orthologous quintet and septet datasets were different from one another, it was unsurprising that there was a lack of concordance observed in the actual genes detected within both analyses. This could be beneficial, in that it acted to enrich the pool of candidate genes for future and subsequent validation analyses. As sequence quality and prediction method improve over time, in addition to the availability of more orthologous sequences, tests of positive selection detection may be improved to provide more reliable sets of genes that have been subject to adaptive evolution on the bovine lineage.

4. Bovine Expressed Sequence Tags (ESTs) As Strategy to Survey Population Diversity

4.1 Introduction

4.1.1 Single Nucleotide Polymorphisms (SNPs)

A single nucleotide polymorphism (SNP) is single-base DNA mutation, and is the most frequently occurring form of genetic variation. SNPs may occur in both coding and non-coding regions of DNA. As they are abundant, and SNP genotyping can be automated at a high-throughput scale, SNPs are used as marker of choice for large scale genome-wide association studies (Hawken et al. 2004) and the inference of population genetics statistics (Akey 2009). A synonymous SNP (synSNP) is a SNP where a point mutation within a codon does not change the corresponding encoded amino acid. In contrast, a non-synonymous SNP (nsSNP) is a point mutation which is amino acid altering, leading to substitution within the protein sequence. A missense SNP is a SNP that results in a premature stop codon, or nonsense codon, that usually results in a truncated protein.

The search for SNPs associated with important traits (e.g. disease) is a major endeavour, and in particular nsSNPs, which are implicated in affecting protein function and fitness, are of interest. nsSNPs have been found to change protein characteristics in the following ways: (a) protein folding and stability, (b) functional sites and biochemical properties, (c) protein expression and localisation, and (d) protein interactions with other proteins or substrates.

Importantly, the effects of nsSNPs can be predicted bioinformatically, based on protein sequence and structural information. SIFT (sorting intolerant from tolerant) is one of several bioinformatics tools available that predict the functional impact of any given SNP (Ng and Henikoff 2003). Alternative available tools include Polymorphism Phenotyping (PolyPhen) (Ramensky, Bork, and Sunyaev 2002), Multivariate Analysis of Protein Polymorphism (MAPP) (Stone and Sidow 2005), and Prediction of Pathological Mutations (PMUT) (Ferrer-Costa et al. 2005).

4.1.2 SNP Discovery Using Expressed Sequence Tag (EST) Alignments

Prior to the Bovine HapMap Project (Gibbs et al. 2009) and the availability of bovine SNP chip, most bovine SNPs were discovered from candidate gene studies and a handful of dedicated SNP discovery projects, such as those by Stone *et al.* (Stone et al. 2002), Heaton *et al.* (Heaton et al. 2002; Heaton et al. 2005) and Werner *et al.* (Werner et al. 2004).

The database of expressed sequence tags (ESTs) has been an invaluable resource for SNP discovery. ESTs are short sequences (200-800 base pairs) usually obtained from single-pass sequencing efforts of complementary DNA (cDNA) libraries. ESTs can be clustered based on overlapping sequence identity between them and the multiple sequence alignments can then be screened for the presence of putative SNPs.

As ESTs are obtained from single-pass sequencing efforts without further validation, the resulting fragments may be of low quality and error-prone, particularly at the ends of the reads, and may not provide complete coverage of the gene sequenced (Aaronson et al. 1996).

The largest EST depository is dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) (Boguski, Lowe, and Tolstoshev 1993; Boguski, Tolstoshev, and Bassett 1994) and it currently (September 2009) contains over 63 million ESTs from nearly 2,000 organisms. The growth of this database has been phenomenal, with the number of ESTs having nearly doubled and the number of organisms having tripled since February 2006 (Nagaraj, Gasser, and Ranganathan 2007). When this project began, there were about 850,000 bovine ESTs and now, there are over 1.5 million.

4.1.3 SIFT: Sorting Intolerant From Tolerant

Disease-causing mutations are more likely to occur at functionally and structurally important sites (Krawczak et al. 2000). SIFT estimates whether a particular SNP causes tolerated or deleterious substitutions and therefore the likelihood of altering the protein function (Ng and Henikoff 2003). SIFT uses a Hidden Markov Model (HMM) on a multiple alignment of homologous protein sequences in order to carry out the prediction.

The steps involved in a SIFT prediction (and as illustrated in Figure 4.1):

1. given a query protein sequence
2. search for similar sequences from a protein database
3. select closely related sequences that may share similar function
4. construct a multiple sequence alignment of the selected sequences
5. calculate normalised probabilities for all possible substitutions at each position
6. apply a cut off value to identify the type of substitutions

The rationale behind this procedure is, given the sequence of a protein, other sequences from the family can be aligned to it to give position-specific information. Conserved residues are more likely to be functionally important, and therefore polymorphisms at one of these evolutionarily conserved positions are postulated to affect the protein function. The procedure also identifies amino acids with similar biochemical properties in order to evaluate the severity of the mutation. Changes between amino acids with similar biochemical properties are more likely to be tolerated. The Venn diagram in Figure 4.2 illustrates the classification of the 20 amino acids according to their properties.

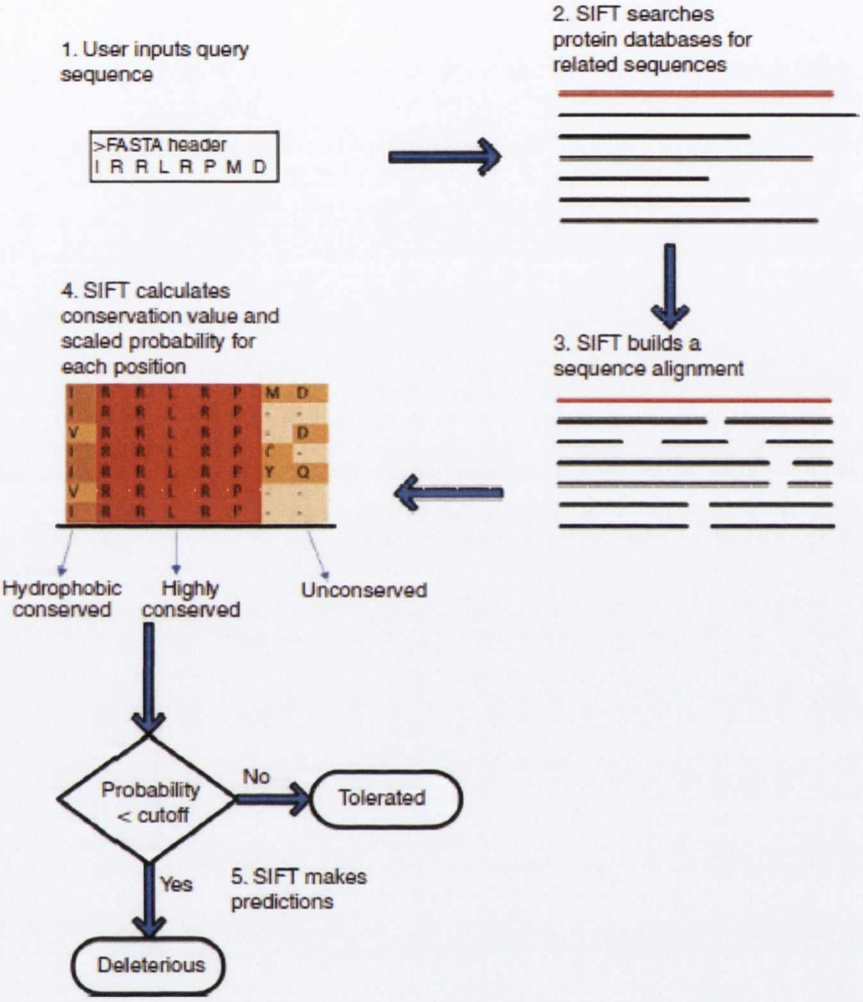


Figure 4.1 | The steps involved in SIFT prediction algorithm (Figure by Kumar *et al.*) (Kumar, Henikoff, and Ng 2009).

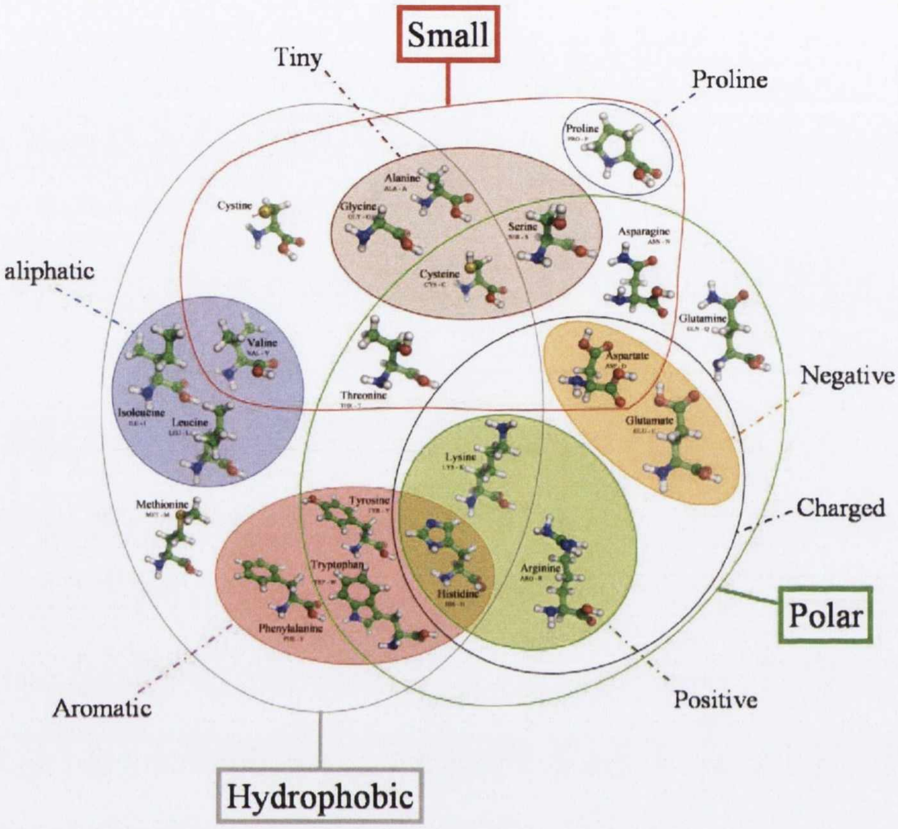


Figure 4.2 | Classical Venn diagram grouping amino acids according to their properties (Image by Alexandre de Brevern, French Institute of Health and Medical Research; redrawn from W. Taylor *et al.*).

4.1.4 Stop Codon Polymorphism

A premature stop codon occurs when a missense point mutation introduces a stop codon within a protein sequence, leading to truncation of the protein. This affects the structure of the protein as well as its biochemical properties and may cause a loss of function of the protein.

In 1999, Olson proposed that loss of function mutations may act as a mechanism to rapidly react to selective pressure (Olson 1999). Loss of function mutations occur more often than function improving mutations (Olson 1999). As this form of mutation is numerous and has an immediate impact on function, the likelihood of selection acting on this class of mutation is also higher. Loss of function mutations may provide such mechanisms in two ways. Firstly, through the direct manifestation of the effect of such mutation, or secondly, through persistence of the mutated gene (if not completely deleted), which will be available for further mutation or reversion when another change in the selective environment occurs. This is in contrast with the evolution of a typical advantageous mutation which requires relatively long evolutionary time to be selected and to become fixed within the population.

There is evidence supporting this theory. Human caspase-12 (CASP12) is polymorphic at amino acid 125, with the ancestral allele encoding the active form of the gene and the derived allele encoding a stop codon leading to a loss of function of the gene (Saleh et al. 2004). This phenotypic change confers an advantage to the individuals carrying this mutation as these individuals have a less pronounced response to bacterial lipopolysaccharide and are therefore less susceptible to sepsis, a sometimes fatal hyper-inflammatory response. This stop codon polymorphism has likely been maintained in populations because of positive selection (Xue et al. 2006).

Another example of beneficial loss of function mutation is seen in the C-C chemokine receptor 5 (CCR5) protein. CCR5 is exploited as an entry point of human immunodeficiency virus (HIV). A deletion of 32 base pairs causes an inactivation of

CCR5. Individuals who are homozygotes for the CCR5-Δ32 mutation have resistance to AIDS and heterozygotes maintain some protection against HIV infection (Dean et al. 1996). While the benefit of this mutation is significant, AIDS is arguably a modern disease for humans. A number of studies in the past had suggested strong positive selection in response to perhaps bubonic plague or smallpox during the middle ages but a recent study by Sabeti *et al.* showed that this mutation may be maintained through neutral evolution (Sabeti et al. 2005).

One example of loss of function mutation in cattle is that of myostatin. Myostatin is a member of the bone morphogenic protein (BMP) family and the tumour growth factor beta (TGF-β) superfamily. It encodes a secreted protein which negatively regulates skeletal muscle growth. A loss of function mutation to myostatin gives rise to the double-muscling phenotype observed in the Belgian Blue and Piedmontese beef cattle breeds (Bellinge et al. 2005). The mutation also leads to very lean meat as it interferes with fat deposition, making them very desirable breeds for quality meat.

As domestication and artificial selection of desirable traits in cattle has exerted severe selective pressure over a short period of time, and given the evidence that stop codon polymorphism has been fundamental in changes that improve the fitness of different organisms, it is possible that certain stop codon polymorphisms have been adopted as one of the potentially adaptive measures to cope with this selective pressure.

In this project, we tried to establish the prevalence of stop codon polymorphisms in cattle genes, validate their presence and investigate the probable effects resulting from such polymorphisms. We also built a database of bovine SNPs in protein coding genes and the predicted effect of the polymorphism.

4.2 Materials and Methods

4.2.1 Bovine EST Assembly

A pipeline was set up to identify SNP sites based on available bovine expressed sequence tags (ESTs) from dbEST. A total of 837,009 and 19,458 ESTs of *Bos taurus* and *Bos indicus* origin, respectively were downloaded. These sequences were vector cleaned using SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) in conjunction with a vector contaminant database – NCBI's UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). The sequences were subsequently masked for repetitive elements and low complexity regions using RepeatMasker (http://www.repeatmasker.org/). The sequences were then collated in FASTA format.

A bioinformatics tool known as TIGR Gene Indices clustering tools (TGICL) (Pertea et al. 2003) was then implemented to cluster the ESTs into contigs. The FASTA sequences were indexed by the algorithm and all-against-all similarity searches were conducted. Certain parameters were applied, including minimum length of sequence overlap (the default is 40 base pairs), minimum percentage identity (the default is 95%) and the maximum mismatch overhang (the default is 30 nucleotides). The overlapping sequences were then processed and clusters were built in an incremental manner. TGICL clustered the ESTs into 59,196 contigs.

4.2.2 SNP Identification

A separate dataset of 33,533 cattle RefSeq coding sequences was downloaded from NCBI, and a reciprocal BlastN (Altschul et al. 1990; Altschul et al. 1997) was carried out against the clustered EST contigs. This resulted in the successful alignment of a total of 10,557 coding sequences to the EST contigs, and enabled identification of the coding regions in the ESTs and provided a reference point of the locations in the gene where polymorphisms were present.

A Perl script was written to identify SNPs present in each gene from the aligned nucleotide sequences. For SNP calling, in order to minimise false calls due to sequencing error, at least two instances of a polymorphism occurring at a particular site across the clustered ESTs were required. The RefSeq sequences were treated as the wild-type/ancestral bases and the new instances of polymorphism as the derived bases.

In order to identify synonymous and nonsynonymous SNPs, the nucleotide sequences were translated into amino acids and SNPs which changed the encoded protein sequence were identified as nsSNPs. Codons which contained termination signals (i.e. TAG, TAA, TGA) were categorised as stop codon polymorphisms.

Nonsynonymous SNPs were further categorised as conservative or non-conservative substitutions, in accordance to the ClustalW (Thompson, Higgins, and Gibson 1994) “strong group” classifications. Non-synonymous amino acid changes that complied with the classification were identified as conservative substitutions whereas those which did not were categorised as non-conservative substitutions.

4.2.3 SIFT Predictions

SIFT v2.1 for Linux was downloaded (http://blocks.fhcrc.org/sift/SIFT.html) and installed on an in-house server. In-house SIFT requires a protein database and the NCBI Genbank protein database Release 154 was downloaded (ftp://ftp.ncbi.nih.gov/genbank/) for this purpose. As SIFT uses Psi-Blast (Altschul et al. 1997) to identify sequences from the database that are similar to the query sequence, the database was first formatted using formatdb.

Two input files for each protein query were used in the analysis. The first was a FASTA formatted sequence file containing each protein query. The second file contained SNPs identified for each gene following the EST clustering pipeline.

SIFT used the FASTA sequence as a query for Psi-Blast against the Genbank database to identify homologous sequences, which were then used to construct a multiple sequence alignment. Based on the alignment, a position-specific probability estimation (PSSM) for each amino acid position was calculated. The PSSM encoded the likelihood of a particular amino acid occurring at a particular position in the alignment. A SNP was predicted to be deleterious if its normalised value was less than the cut off (the default was 2.75). If a SNP had a PSSM greater than or equal to the cut off, it was predicted to be tolerated.

Figure 4.3 illustrates the pipeline involved, from the clustering of ESTs to identification and classification of SNPs, and subsequent SIFT analysis.

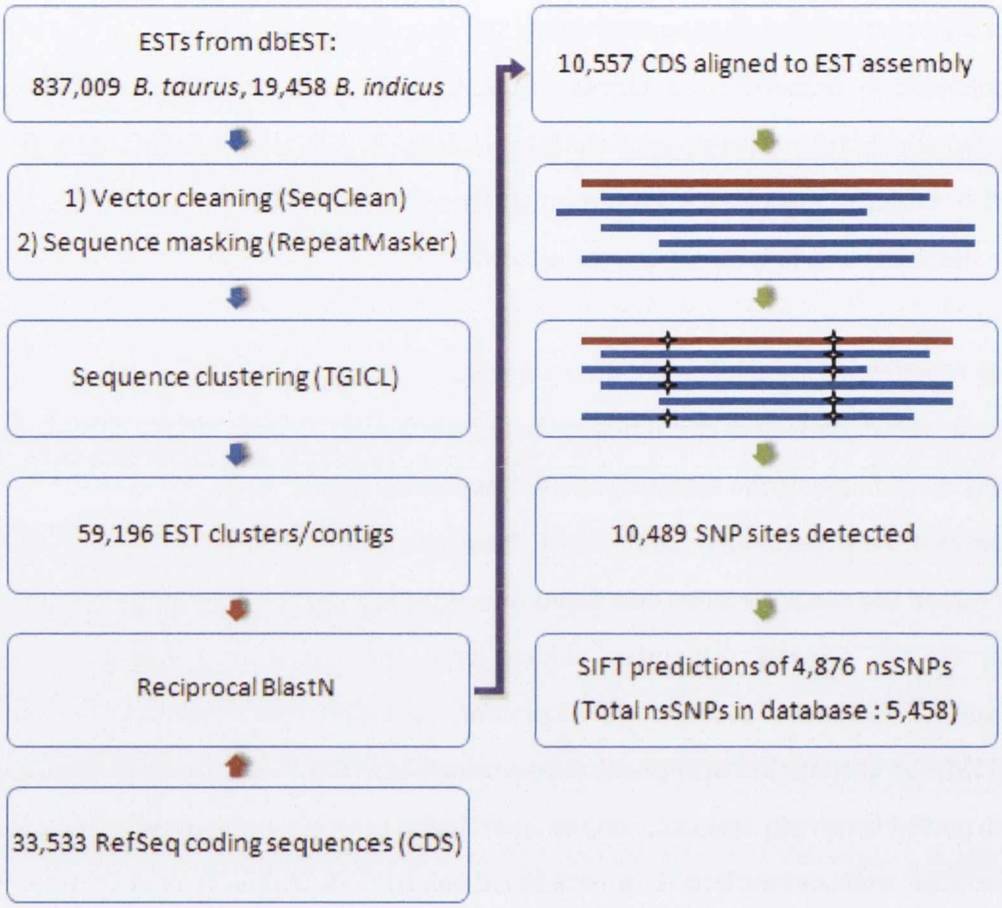

ESTs from dbEST:
837,009 *B. taurus*, 19,458 *B. indicus*
1) Vector cleaning (SeqClean)
2) Sequence masking (RepeatMasker)
Sequence clustering (TGICL)
59,196 EST clusters/contigs
Reciprocal BlastN
33,533 RefSeq coding sequences (CDS)
10,557 CDS aligned to EST assembly
10,489 SNP sites detected
SIFT predictions of 4,876 nsSNPs
(Total nsSNPs in database : 5,458)

Figure 4.3 | Analysis pipeline for clustering of bovine ESTs, alignment to reference sequences, identification of SNPs and subsequent SIFT analysis.

4.2.4 Bovine EST Browser

A web-based browser (internal use only) was created to allow user-friendly searching and visualisation of EST-identified SNPs in each contig; in addition to the SIFT prediction for each SNP, and the associated alignments to access alignment quality. The browser also incorporated information from Panther Ontology (Mi et al. 2005) to provide insight into the biological processes and molecular functions of genes that contain the identified SNPs. The browser also allows querying for genes which contain potential stop codon polymorphisms.

4.2.5 Quality Assessment of Stop Codon Polymorphisms

For any genes that were identified as putatively containing stop codon polymorphisms, manual assessments were carried out to establish the support for such mutation. The quality of the multiple sequence alignments were checked, the presence of other SNPs in the region were investigated to identify if there were particular splice variants that might account for the stop codon in some sequences. The locations of the SNPs in relation to the EST sequences were also identified to determine if the SNP was likely to be sequencing error, particularly for the regions of end sequences. Additionally, trace files of the corresponding ESTs used in the alignments to identify stop codon polymorphism that were submitted to dbEST were checked to verify that the nucleotide calls had a suitable quality score. Trace files are essentially sequencing data of a contiguous DNA fragment in one orientation of sequencing (either forward or reverse strand sequencing) and contain intensity measurements of the four genetic bases. When a base has the strongest emission intensity, the nucleotide is called as part of the gene sequence. There are two main public libraries of trace archive that work hand-in-hand to coordinate the file storage – one maintained by NCBI (http://www.ncbi.nlm.nih.gov/Traces/home/) known simply as Trace Archives, and another by Ensembl known as Ensemble Trace Server (http://trace.ensembl.org/).

4.2.6 Resequencing in Populations

Candidate gene(s) that were predicted to putatively contain a stop codon polymorphism were judged to be candidates for sequencing in different cattle populations in order to validate the presence of the stop codon, and to carry out additional population genetic analysis.

A panel of cattle samples, consisting of 31 European *Bos taurus* (16 Aberdeen Angus and 15 Friesian), 16 African *Bos taurus* (Lagune), 7 Asian *Bos indicus* (2 Hariana, 2 Tharparkar, 2 Sahiwal, 1 Ongole) and 1 outgroup (plains bison), were used for population resequencing. Extracted DNA samples were available in the laboratory from previously published studies (Tables 4.1).

Table 4.1 | Breed information of bovine samples in the panel used for EST SNP sequencing.

Breed	Origin	N	Sample Ids
Friesian	Europe	15	FR19-FR20, FR26-FR34, FR37-FR40
Aberdeen Angus	Europe	16	AA1-4, AA10-AA21
Lagune	Africa	16	L20-L34, L50
Hariana	Asia	2	Har5a, Har10a
Tharparkar	Asia	2	Thr3b, Thr10b
Sahiwal	Asia	2	Sah6a, Sah10a
Ongole	Asia	1	Ong26
Plains Bison		1	PB

4.2.7 Primer Design, Polymerase Chain Reaction (PCR) and Sequencing

Primers were designed using Genefisher (http://bibiserv.techfak.uni-bielefeld.de/genefisher2/submission.html) and ordered from VH Bio Ltd, England (http://www.vhbio.com). They were designed to be between 18 to 24 base pairs in length, containing 45-55% GC content and have annealing temperate between 55°C to 65°C.

Initial optimisation of PCR conditions was carried out on the MJ research DYAD PCR machine. A temperature gradient between 60°C to 64°C was investigated to identify the best annealing temperature, and at each temperature gradient, $MgCl_2$ concentrations of 1.5μM and 2.0μM were tested.

PCR was carried out in 96 well plates, with a total reaction volume of 15μL per sample. The samples were sent to AGOWA, Germany for sequencing. Two amplifications were carried out for each sample, one utilising the forward primer and one utilising the reverse primer.

The reagent concentrations in the reaction volume for the PCR were: 1x PCR buffer, 0.2μM of each dNTP, 0.5U/μL of Taq polymerase. The cycling conditions for amplifications were: 95°C for 15 minutes, followed by 30 cycles of 95°C for 40 seconds, annealing temperature (T_a) for 40 seconds and 72°C for 1 minute, followed by a final extension step of 72°C for 10 minutes.

The experimental work of this section was completed with the assistance of Dr. Valeria Mattiangeli.

4.3 Results

4.3.1 Bovine EST Assembly and SNP Statistics

A total of 837,009 *Bos taurus* and 19,458 *Bos indicus* ESTs were obtained, vector-cleaned, sequence-masked and clustered into 59,196 contigs. Following a BlastN search, 10,557 of these contigs were successfully aligned to RefSeq coding sequences (CDS). SNPs were identified from the RefSeq CDS-EST contig alignments, using the bases in the CDS as the major allele. 10,489 SNP sites were detected. Our interest primarily lay in the detection of SNPs that are potentially function-altering, therefore SNPs that have introduced non-synonymous or stop mutations were focused on.

Among the 10,489 SNPs, 5,458 of them were identified as non-synonymous SNPs (nsSNPs) of which 376 had no SIFT prediction information and 59 were not scored by SIFT. A further subset of 147 SNPs contained putative stop codon polymorphisms. Of the 4,876 nsSNPs with SIFT predictions, 2,469 nsSNPs were predicted to be "tolerated" substitutions while 2,407 nsSNPs were predicted to be "deleterious". In the "tolerated" dataset, 606 were annotated to have undergone a non-conservative substitution whereas 1,863 were predicted as conservative substitutions. Within the "deleterious" mutation dataset, 1,247 were non-conservative substitutions and 1,160 were conservative substitutions. Figure 4.4 summarises these numbers.

The remainder 5,031 SNPs were synonymous SNPs (synSNPs) and have no further SIFT prediction, given that synSNPs are non-amino acid altering.

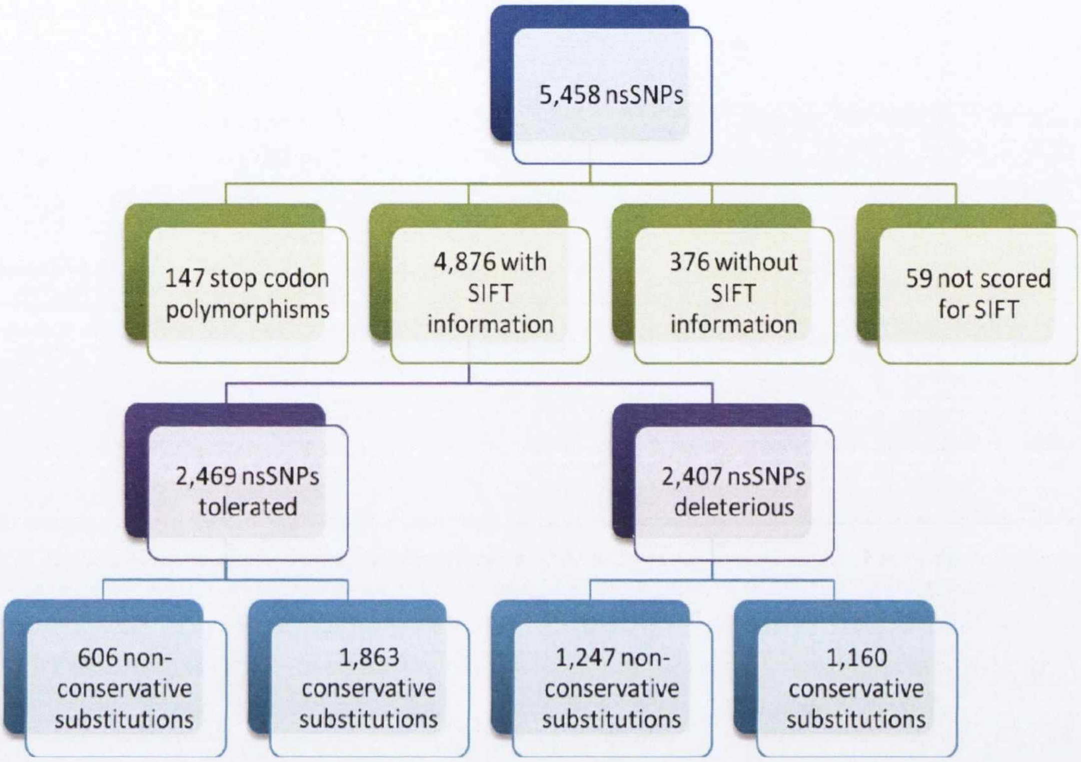

5,458 nsSNPs
147 stop codon polymorphisms
4,876 with SIFT information
376 without SIFT information
59 not scored for SIFT
2,469 nsSNPs tolerated
2,407 nsSNPs deleterious
606 non-conservative substitutions
1,863 conservative substitutions
1,247 non-conservative substitutions
1,160 conservative substitutions

Figure 4.4 | Summary of the results of the SIFT analysis of nsSNPs in bovine ESTs. The numbers of predicted tolerated and deleterious alleles are shown.

The focus of this study was on nsSNPs (particularly deleterious SNPs) and stop codon polymorphisms. Gene Ontology was used to characterise nsSNP containing genes in terms of their biological processes and molecular functions to investigate whether any particular category was over-represented in deleterious nsSNPs.

InnateDB (http://www.innatedb.com), a platform that facilitates system analyses using integrated pathway and molecular interaction data from publicly available databases (Lynn et al. 2008a), was used to investigate which of the ontology groups linked to nsSNPs were over-represented. A list of bovine genes with predicted nsSNPs were first assigned to their human orthologues, and then uploaded for gene ontology over-representation analysis (Ontology ORA).

Over-representation was found in genes which contain nsSNPs involved in a total of 40 biological process categories that include cell cycle, mitosis, translation, translational initiation, translational elongation, protein folding, protein transport, intracellular protein transport, intracellular protein transmembrane transport, membrane organisation and metabolic process.

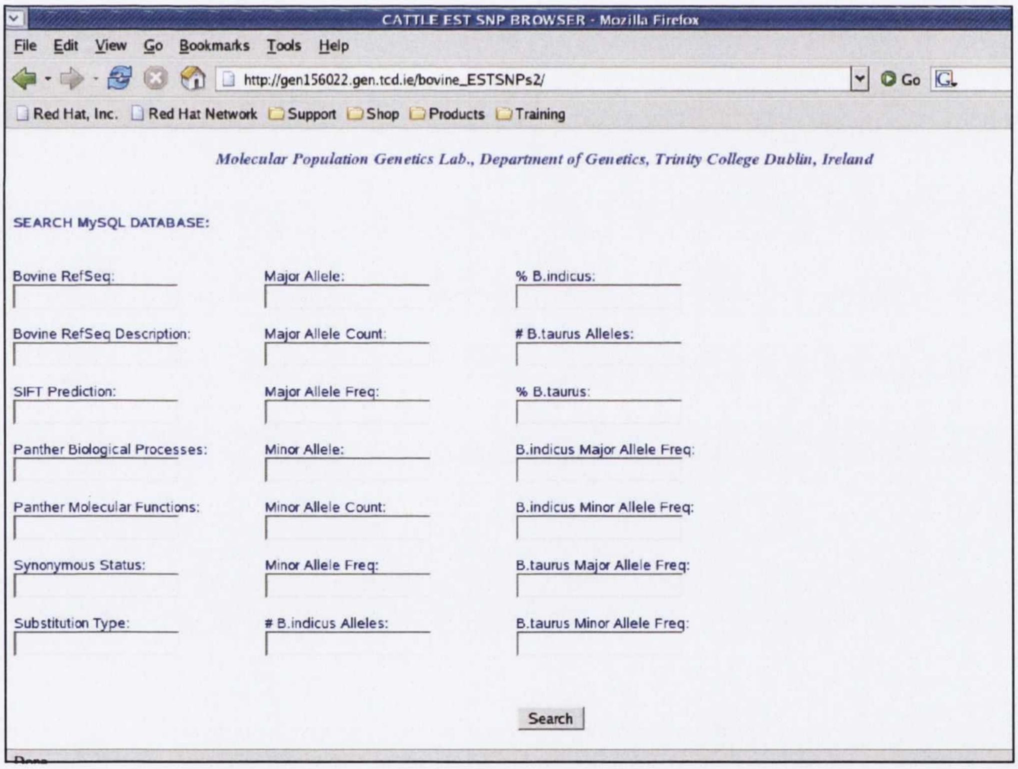
In terms of molecular functions, there were 23 over-represented categories including antioxidant activity, catalytic activity, lyase activity, structural constituent of ribosome, translation initiation factor activity, protein binding, RNA binding, rRNA binding, lipid binding and fatty acid binding.

See Supplementary Table 4.1 for the complete output of the InnateDB Ontology ORA analyses of nsSNP-containing genes.

4.3.2 Bovine EST Browser

Following the EST assembly pipeline, a MySQL database was set up containing annotation associated to each of the genes analysed, including RefSeq accession number, gene description, SNP position, SNP count, type of substitution and SIFT analysis result.

A simple bovine EST browser (internal use only; Figure 4.5) was then built to facilitate queries of the information within the database using an input form. The form was written in HTML, and PHP was used to send queries to the MySQL database set up earlier.


CATTLE EST SNP BROWSER - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://gen156022.gen.tcd.ie/bovine_ESTSNPs2/
Go
Red Hat, Inc.
Red Hat Network
Support
Shop
Products
Training
Search

Molecular Population Genetics Lab., Department of Genetics, Trinity College Dublin, Ireland

SEARCH MySQL DATABASE:

Bovine RefSeq:

Bovine RefSeq Description:

SIFT Prediction:

Panther Biological Processes:

Panther Molecular Functions:

Synonymous Status:

Substitution Type:

Major Allele:

Major Allele Count:

Major Allele Freq:

Minor Allele:

Minor Allele Count:

Minor Allele Freq:

B.indicus Alleles:

% B.indicus:

# B.taurus Alleles:

% B.taurus:

B.indicus Major Allele Freq:

B.indicus Minor Allele Freq:

B.taurus Major Allele Freq:

B.taurus Minor Allele Freq:

Figure 4.5 | Screenshot of bovine EST browser (for internal use only).

The inputs for searching are straight forward, and the query can be made based on a single criterion or a combination of them. The input fields take the following formats for query:

- Bovine RefSeq: RefSeq accession number e.g. NM_001007813
- Bovine RefSeq Description: description of a gene e.g. tumor necrosis factor
- SIFT Prediction: either tolerated, deleterious or no SIFT
- Synonymous Status: either "synSNP" or "nsSNP"
- Substitution Type: either "synonymous SNP", "conservative substitution" or "non-conservative substitution"
- Major/Minor Allele: single letter amino acid code, or STOP for stop codon
- Major/Minor Allele Count: integer from 1 onwards
- Major/Minor Allele Frequency: decimal between 0 and 1
- B. taurus/ B. indicus Allele Count: integer from 1 onwards
- B. taurus/ B. indicus Allele Frequency: decimal between 0 and 1
- B. taurus/ B. indicus Percentage: number between 0 and 100
- Panther Biological Processes: terms of biological processes, e.g. cell structure, mRNA splicing, cell adhesion-mediated signalling, immunity and defense
- Panther Molecular Functions: terms of molecular functions, e.g. ion channel, RNA helicase, extracellular matrix

4.3.3 Verification of Stop Codon Polymorphisms

The database contains 147 predicted stop codon polymorphisms. The EST browser was used to query for genes containing stop codon polymorphisms, with a minor allele frequency of at least 0.15 (the range of MAF frequently used is between 0.15 and 0.25, where MAF less than 0.10 is normally considered too low, while 0.25 is the typical frequency value of common minor allele) and a minor allele count of at least three (this is to ascertain that the polymorphism called is more likely to be real than being a sequencing error). A total of 14 hits were found, which were then subjected to a series of manual assessments to determine their likelihood of being true positives. One of the genes was eliminated as it was removed from the NCBI database as a result of standard genome annotation processing.

First, the sequence alignments were checked to determine that they were of high quality. The locations where the SNPs were positioned were noted, and this is of importance because it is a known issue that end sequences are more error-prone while regions in the middle of reads are of higher quality. Other SNPs that were found near the stop codon polymorphisms were also taken into account, as this may be indicative of the presence of an alternative splice variant which could account for the stop codon (i.e. a shorter splice variant with 3' UTR could appear to have a stop codon in comparison with the RefSeq sequence).

The EST sequences used to build the alignments were also cross-checked against the trace archive to ensure that the base calls were unambiguous and therefore accurate.

The verification of the evidence to support/reject the presence of stop codon polymorphisms is presented (Table 4.2) and the candidate most likely to truly contain a novel stop codon polymorphism was then resequenced in a panel of population samples in an attempt to verify it.

Table 4.2 | Stop codon polymorphisms and outcome of manual verifications.

RefSeq ID	Gene Symbol	SNP Position	Note / Comment
NM_001038086	TCP1	R 526 Stop	No trace match found
XM_591518	AKAP12	E 1549 Stop	No trace match found
XM_588536	LOC511239	E 129 Stop	Trace not unambiguous for support
NM_001034453	CCDC104	W 11 Stop	Start of EST sequences, likely to be error
XM_583037	RBP4	G 139 Stop	Start of EST sequences, likely to be error
XM_867922	ZFN1	H 168 Stop	Probable end of an alternative transcript
XM_865941	DERL3	W 156 Stop	Probable end of an alternative transcript
NM_001015621	EIF2S2	Q 40 Stop	Multiple SNPs, likely alternative splice site
NM_001034677	GNAI3	A 30 Stop K 35 Stop	Multiple SNPs, likely alternative splice site
XM_601113	G6PD	A 30 Stop	Multiple SNPs, likely alternative splice site
XM_614568	TLK2	K 961 Stop	Multiple SNPs, likely alternative splice site
XM_616549	LOC536418	K 115 Stop	Good traces, but gene removed from NCBI
XM_864316	SOCS1	E 201 Stop	Good traces found, good alignment

4.3.4 Suppressor of Cytokine Signalling 1 (SOCS1)

Following the *in silico* stop codon polymorphism prediction and verification of the alignment and base-call qualities, suppressor of cytokine signalling 1 (SOCS1) was found to be the strongest candidate containing a putative stop codon polymorphism. The quality of the multiple sequence alignment of the ESTs was high, with 3 out of 7 of the EST sequences containing the stop codon polymorphism (Figure 4.6). Additionally, good supporting traces of the ESTs were also found.

The SOCS1 gene was discovered in 1997 to be involved in a JAK/STAT negative feedback loop to attenuate cytokine signalling (Endo et al. 1997). It has been shown to inhibit signalling of a wide range of cytokines including interleukin-2 (IL2) (Sporri et al. 2001), IL4 (Losman et al. 1999), IL6 (Endo et al. 1997) and prolactin (Pezet et al. 1999). This is interesting to us, particularly as SOCS1 has a role not only in immunity but also in lactation, both of which would have been the source of significant selective pressure following cattle domestication.

The SOCS1 gene contains two conserved domains: (a) SH2 domain and (b) the SOCS box (Figure 4.7). The central SH2 domain is approximately 95 amino acids in length and has been shown to be essential for binding to kinase receptor domain in order to initiate the cytokine signalling cascade. The conserved 40-residue SOCS box is found at the C-terminal and is essential for association with Elongin B/C complex. This complex has been postulated to be involved in stabilisation of SOCS1.

The predicted stop codon polymorphism was located within the conserved SOCS box. The C/T polymorphism changes residue 201 in the protein from glutamic acid to a premature stop codon. We hypothesised that this stop codon polymorphism may be a mechanism to increase cytokine signalling in the immune response or potentially to improve lactation in cattle.

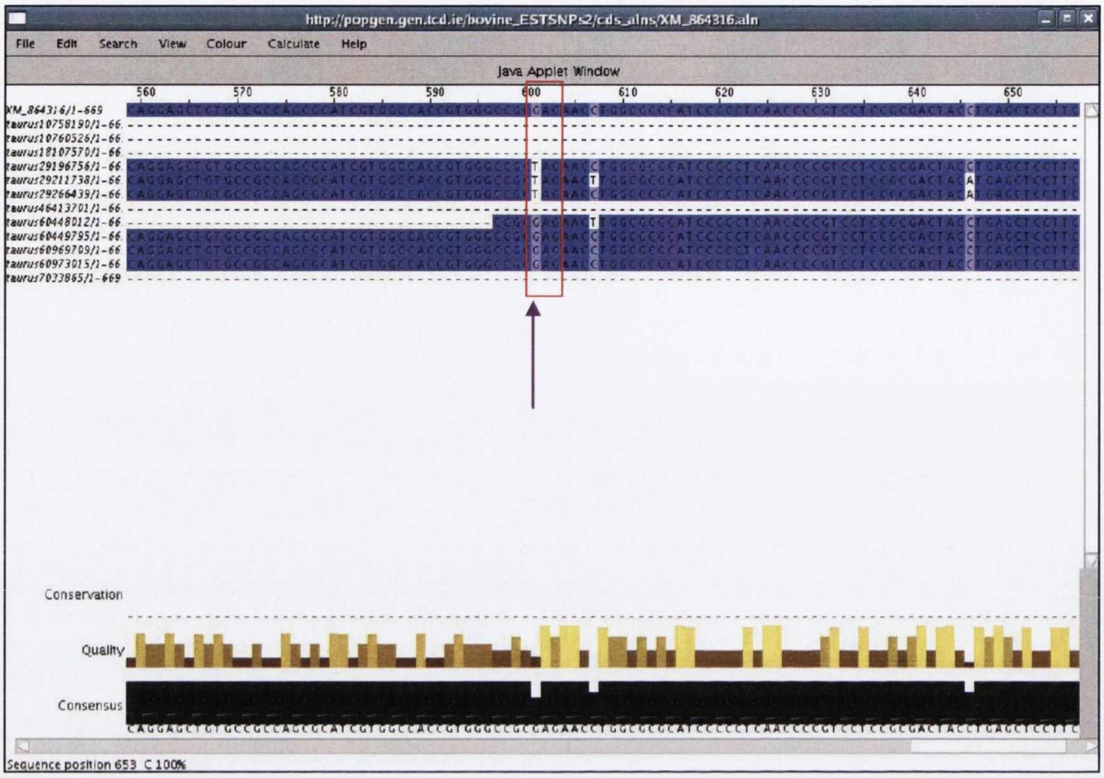


Figure 4.6 | Multiple alignment of genomic sequence (i.e. the first sequence) and the ESTs to SOCS1, showing sequences of flanking region to the stop codon. The position of the stop codon is marked by the red box and the position where polymorphism occurs is marked by the arrow.

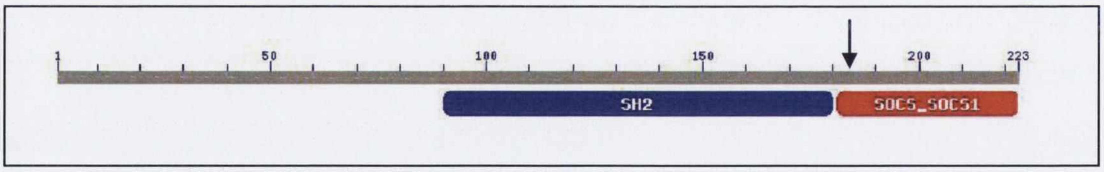


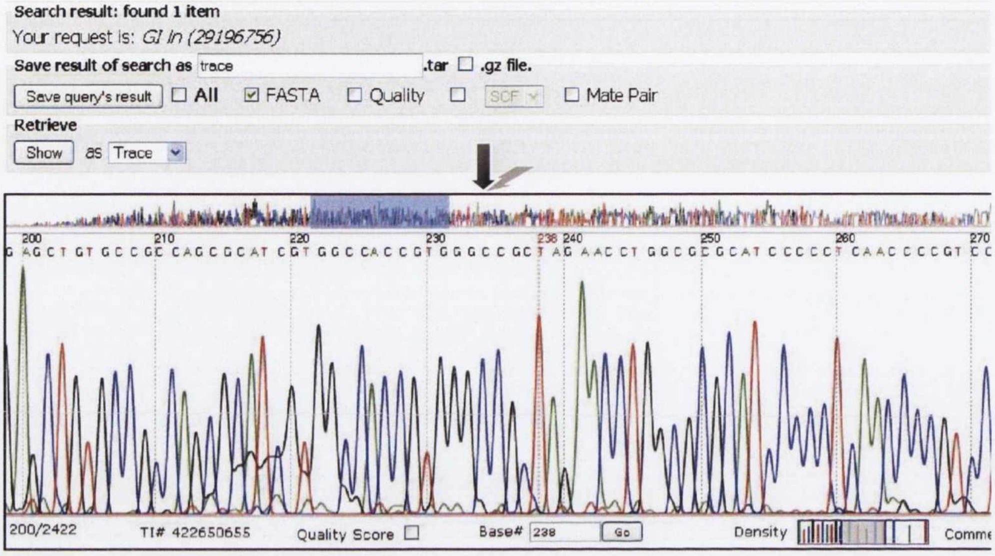
Figure 4.7 | Graphical representation of SOCS1, showing the conserved SH2 and SOCS box domains. The location of the putative stop codon is marked by the arrow.

Good chromatogram trace support was found for the stop codon polymorphism-containing EST sequences in the alignment, each with distinct trace peaks to verify that these are not instances of unreliable base calls. An additional EST purporting to carry this same mutation was also found through BLAST search of the trace archive. (Table 4.3; Figures 4.8-4.11)

Table 4.3 | Trace support of stop codon polymorphism-containing ESTs and locations of stop codon polymorphism based on EST nucleotide sequences.

EST	**EST**	**Position of**	
GI Accession	**Trace Index**	**Stop Codon**	**Trace Orientation**
29196756	422650655	238	Forward
29211738	422658164	568	Reverse
29266439	425684720	560	Reverse
29264544	425682827	258	Forward

Figures 4.8-4.11 | Trace support for stop codon polymorphism for ESTs listed in Table 4.4, with arrows to indicate the locations where polymorphisms occur.



Search result: found 1 item

Your request is: *GI in (29196756)*

Save result of search as trace **.tar** □ **.gz file.**

Save query's result □ **All** ☒ FASTA □ Quality □ SCF □ Mate Pair

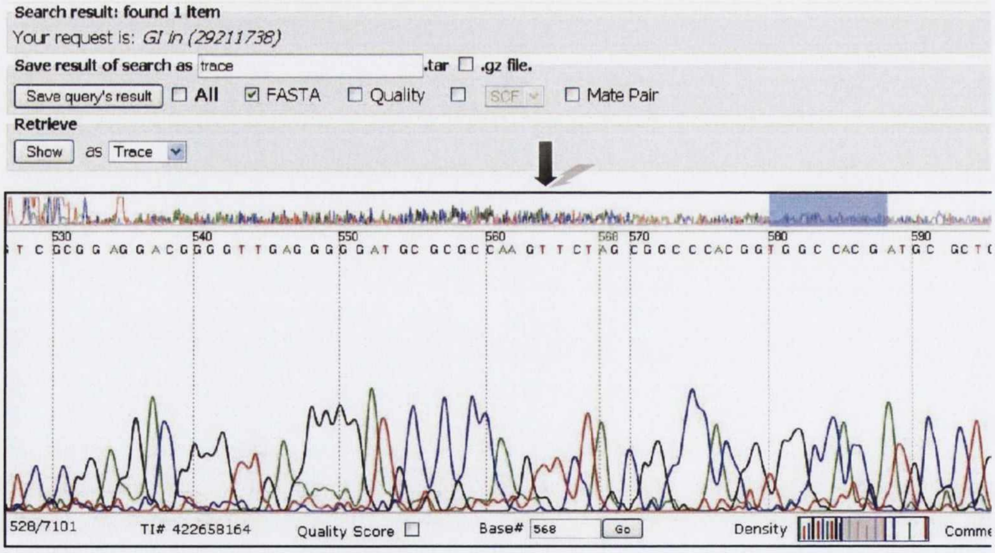
Retrieve

Show as Trace

200/2422 TI# 422650655 Quality Score □ Base# 238 Go Density Comme

200
210
220
230
238 240
250
260
270
G AGCT GT GCCGC CAGCGCAT CGT GGCCACCGT GGGCCGCT AG AACCT GGCGC GCAT CCCCCT CAACCCCGT CC

Figure 4.8 | Trace for EST GI 29196756.



Search result: found 1 item

Your request is: *GI in (29211738)*

Save result of search as trace **.tar** □ **.gz file.**

Save query's result □ **All** ☒ FASTA □ Quality □ SCF □ Mate Pair

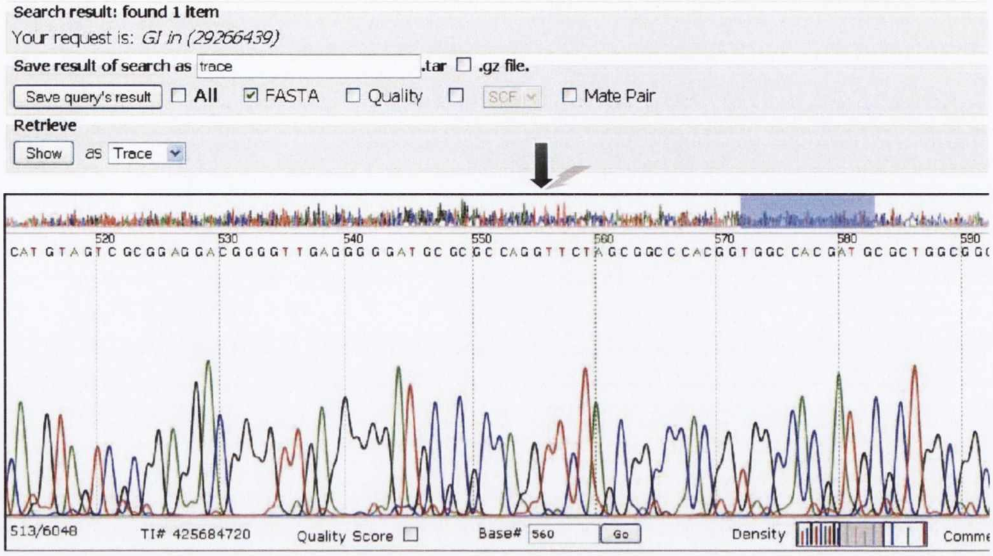
Retrieve

Show as Trace

528/7101 TI# 422658164 Quality Score □ Base# 568 Go Density Comme

530
540
550
560
568 570
580
590
T C GCGG AGGACGGG GTTGAG GGG GAT GC GCGC CAA GT T CT AG CGGCCCACGGT GGC CAC GAT GC GCT C

Figure 4.9 | Trace for EST GI 29211738.



Search result: found 1 item

Your request is: *GI in (29266439)*

Save result of search as trace **.tar** □ **.gz file.**

Save query's result □ **All** ☑ FASTA □ Quality □ SCF □ Mate Pair

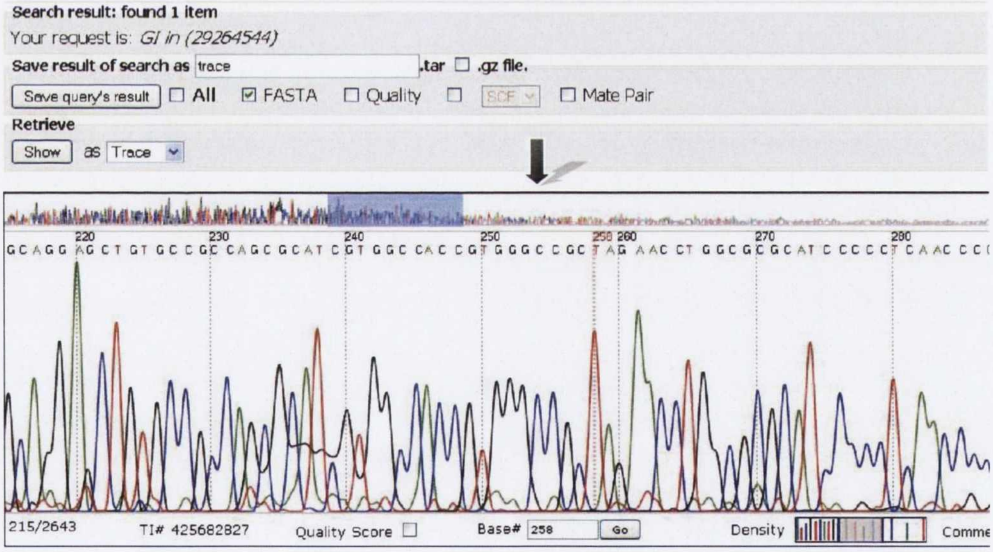
Retrieve

Show as Trace

513/6048 TI# 425684720 Quality Score □ Base# 560 Go Density Comme

520
530
540
550
560
570
580
590
CAT GTAGT C GC GG AG GAC GGGGT T GAG GG G GAT GC GC GC CAGGT T CTA GC GGC C CAC GGT GGC CAC GAT GC GC T GGC G GC

Figure 4.10 | Trace for EST GI 29266439.



Search result: found 1 item

Your request is: *GI in (29264544)*

Save result of search as trace **.tar** □ **.gz file.**

Save query's result □ **All** ☑ FASTA □ Quality □ SCF □ Mate Pair

Retrieve

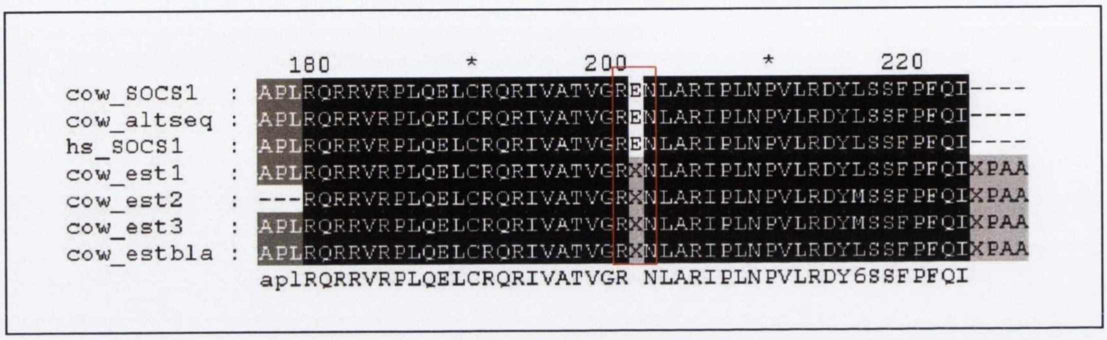
Show as Trace

215/2643 TI# 425682827 Quality Score □ Base# 258 Go Density Comme

220
230
240
250
258 260
270
280
GCAGG AG CT GT GC C GC CAGC GCAT C GT GGC CAC CG T GGGC C GC T AG AAC CT GGC G C GCAT C CCCC T CAAC CC

Figure 4.11 | Trace for EST GI 29264544.

The SOCS1 reference coding sequences as well as the EST coding sequences containing the stop codon polymorphism were obtained, translated and aligned. A human SOCS1 was also added to show sequence conservation for the protein (Figure 4.12).



```
                 180         *       200         *       220
cow_SOCS1   : APLRQRRVRPLQELCRQRIVATVGRENLARIPLNPVLRDYLSSFPFQI----
cow_altseq  : APLRQRRVRPLQELCRQRIVATVGRENLARIPLNPVLRDYLSSFPFQI----
hs_SOCS1    : APLRQRRVRPLQELCRQRIVATVGRENLARIPLNPVLRDYLSSFPFQI----
cow_est1    : APLRQRRVRPLQELCRQRIVATVGRXNLARIPLNPVLRDYLSSFPFQIXPAA
cow_est2    : ---RQRRVRPLQELCRQRIVATVGRXNLARIPLNPVLRDYMSSFPFQIXPAA
cow_est3    : APLRQRRVRPLQELCRQRIVATVGRXNLARIPLNPVLRDYMSSFPFQIXPAA
cow_estbla  : APLRQRRVRPLQELCRQRIVATVGRXNLARIPLNPVLRDYLSSFPFQIXPAA
              aplRQRRVRPLQELCRQRIVATVGR NLARIPLNPVLRDY6SSFPFQI
```

Figure 4.12 | Alignment of SOCS1 reference bovine and human protein sequences, and translated bovine EST of the four stop-codon containing sequences shown in Figures 4.8 to 4.11. The stop codon gave rise to "X" contained within the red box.

4.3.5 Resequencing of SOCS1

The preliminary analyses showed SOCS1 to be a strong candidate to investigate for stop codon polymorphism. SOCS1 is a gene of immunological importance, and therefore is a possible target for disease-related selection. A stop codon polymorphism is a dramatic mutation that could form a focus for such selection. Therefore it was decided to validate the presence of this mutation in a panel of test samples.

A set of primers of 20 bases each were designed for sequencing of the region around this SNP in different cattle populations.

Forward primer 5'-3': AGC CGC GAG AGC TTC GAC TG

Reverse primer 5'-3': GAG GGC GCC CCA GTT AAT GC

The optimised PCR conditions for these primers are T_a of 61°C at $MgCl_2$ concentration of 1.5μM.

Not all samples were successfully sequenced. Of 55 samples tested, sequencing of 11 samples using both forward and reverse primers were unsuccessful – FR19, FR20, FR26, FR29, FR31, FR32, AA2, AA3, AA15, AA16 and PB. A further 3 samples could not be sequenced based on the forward primer – FR33, AA18 and L21.

The sequences were aligned using software from Molecular Evolutionary Genetic Analysis (MEGA, http://www.megasoftware.net/) (Tamura et al. 2007) and visualised and edited in GeneDoc (http://www.nrbsc.org/gfx/genedoc/index.html). The multiple alignments of all sequenced nucleotides and of only the SOCS box, based on forward and reverse sequencing, are displayed in Figures 4.13 to 4.16.

The results showed that, for both the forward and the reverse sequencing strands, all sequences contain the nucleotide bases that reflect the reference sequence, and that polymorphism was completely absent at the position where the stop codon was thought to be present.

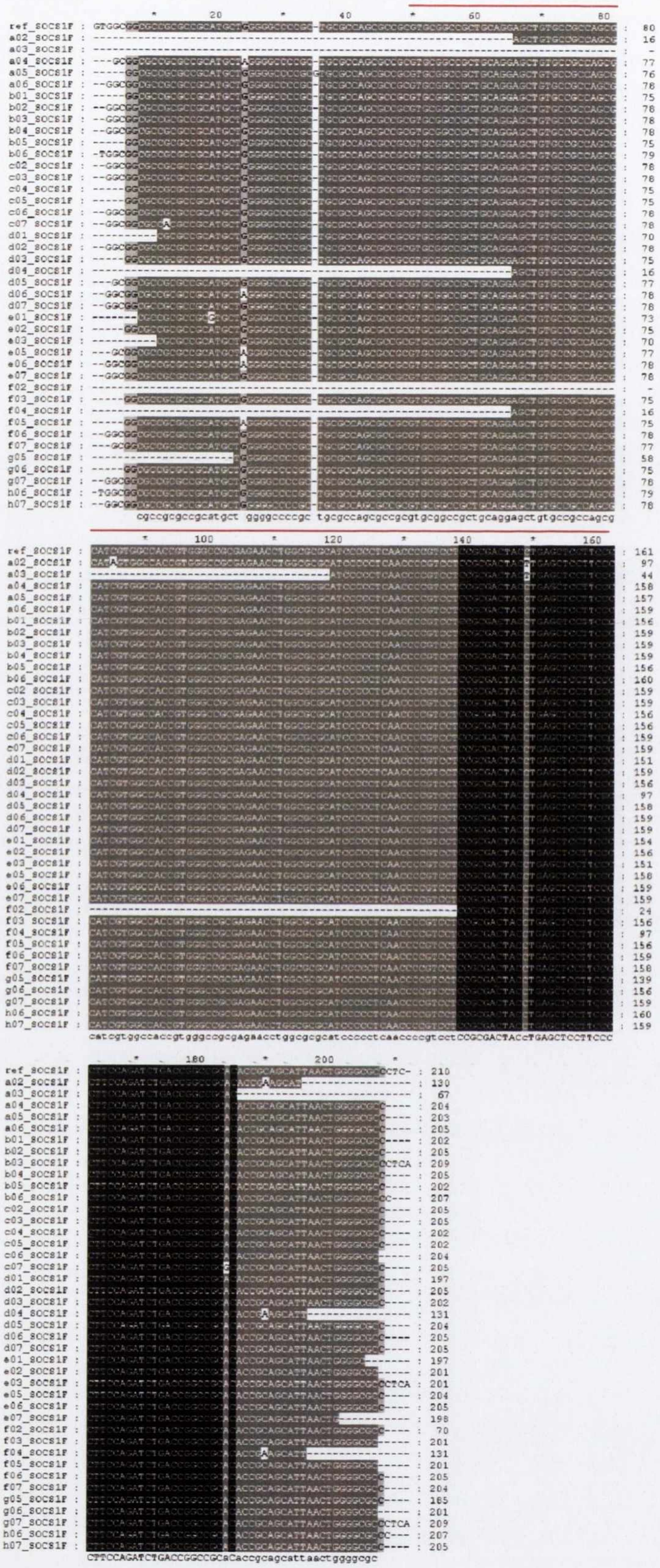


Figure 4.13 | Multiple alignment of sequences around SOCS box based on forward-strand sequencing. SOCS box domain runs along the positions marked by the red line.

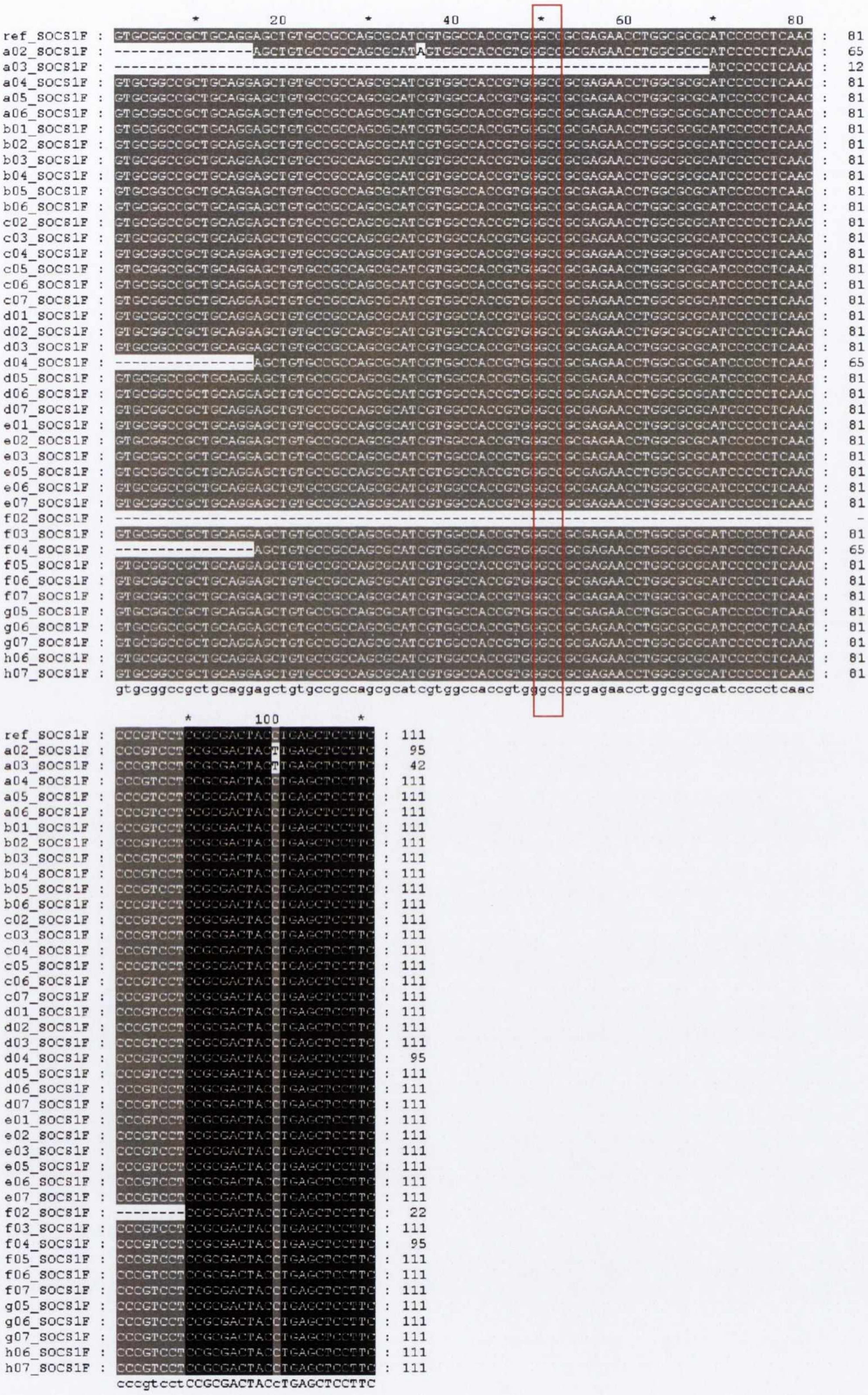
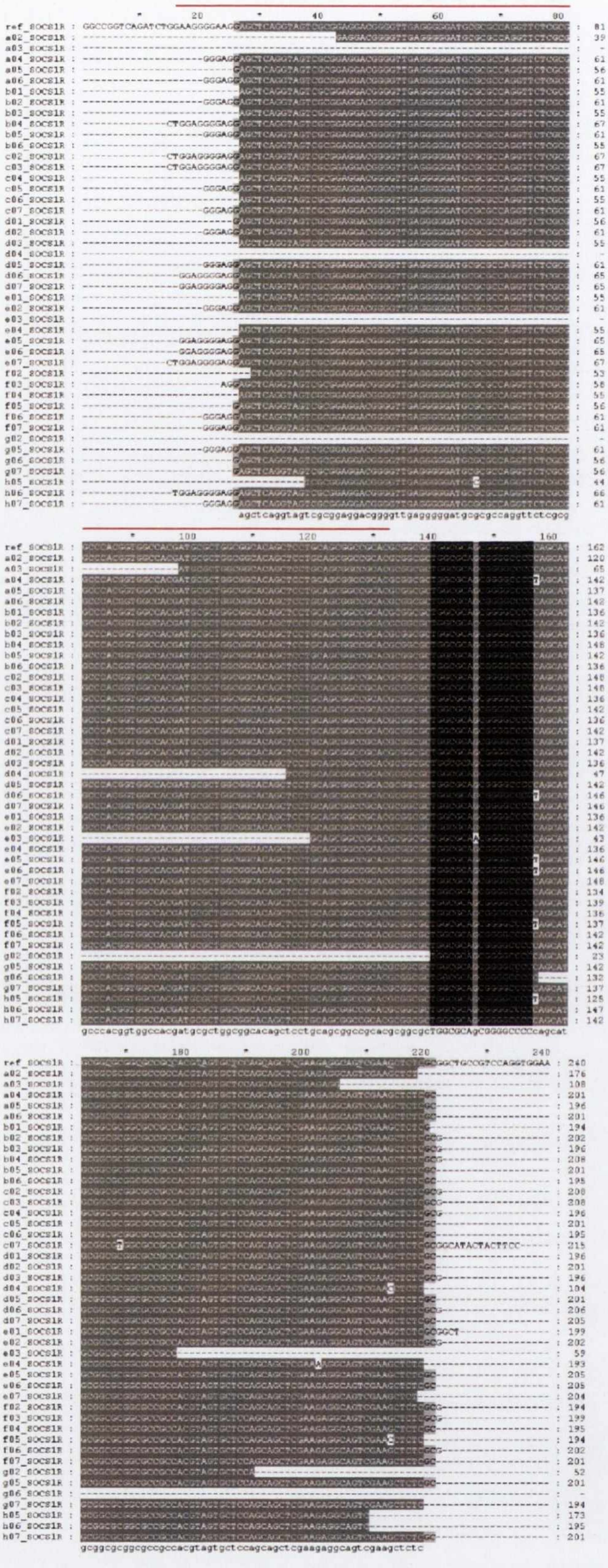
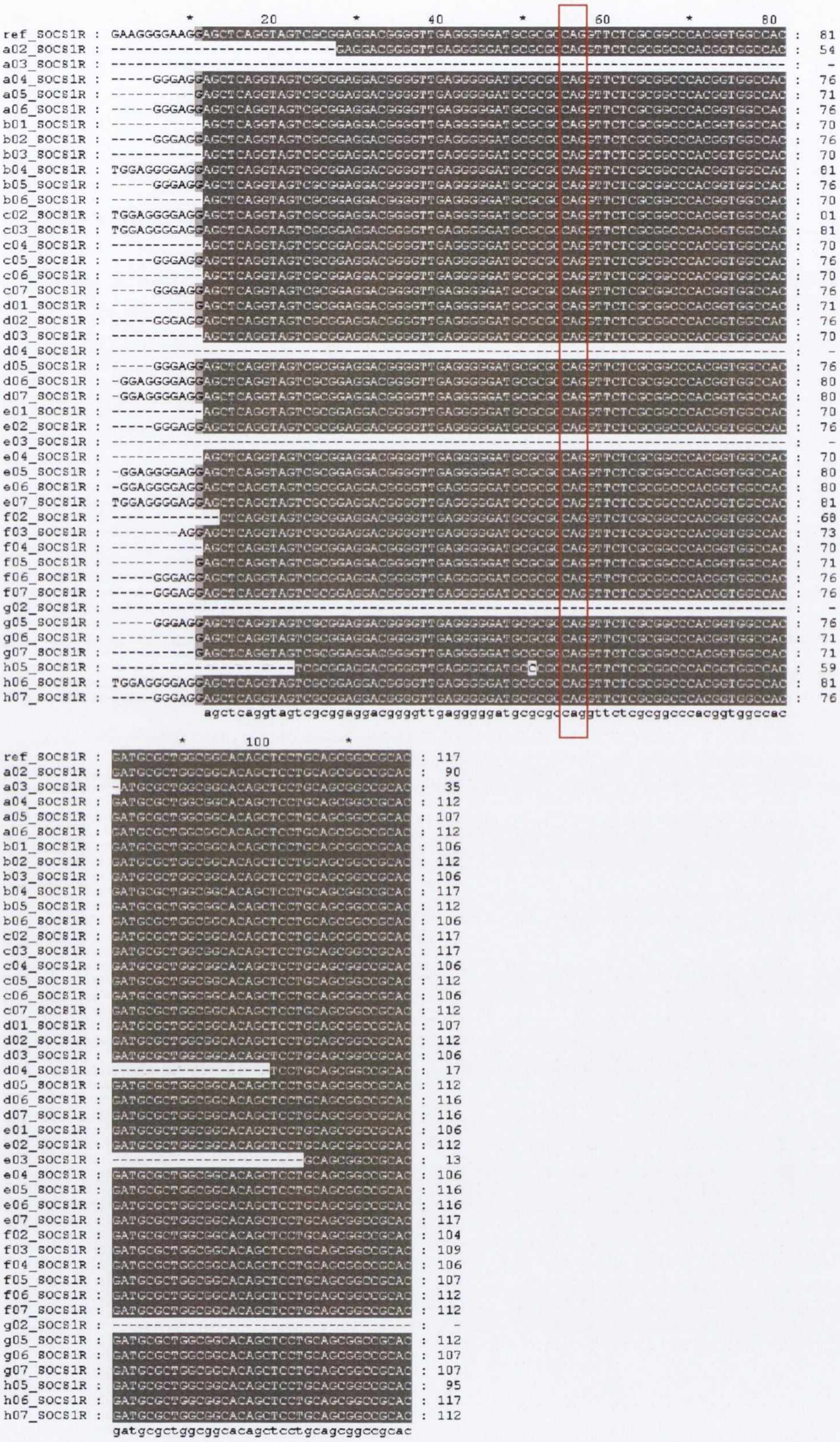


Figure 4.14 | Multiple alignment of SOCS box based on forward sequencing. The predicted position of stop codon polymorphism is highlighted by the red box.



20
40
60
80
100
120
140
160
180
200
220
240
ref_SOCS1R
agctcaggtagtcgcggaggacggggttgaggggatgcgcgccaggttctcgcg
gcccacggtggcccacgatgcgctggcggcacagctcctgcagcggccgcacgcggcgctGGCGCAGCGGGGCCCCcagcat
gcggcgcggcgccgccacgtagtgctccagcagctcgaagaggcagtcgaagctctc

Figure 4.15 | Multiple alignment of sequences around SOCS box based on reverse-strand sequencing. SOCS box domain runs along the positions marked by the red line.



```
                        *        20         *        40         *        60         *        80
ref_SOCS1R : GAAGGGGAAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  81
a02_SOCS1R : ---------------------------GAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  54
a03_SOCS1R : --------------------------------------------------------------------------------- :   -
a04_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
a05_SOCS1R : ----------GAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  71
a06_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
b01_SOCS1R : -----------AGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  70
b02_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
b03_SOCS1R : -----------AGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  70
b04_SOCS1R : TGGAGGGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  81
b05_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
b06_SOCS1R : -----------AGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  70
c02_SOCS1R : TGGAGGGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  81
c03_SOCS1R : TGGAGGGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  81
c04_SOCS1R : -----------AGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  70
c05_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
c06_SOCS1R : -----------AGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  70
c07_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
d01_SOCS1R : ----------GAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  71
d02_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
d03_SOCS1R : -----------AGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  70
d04_SOCS1R : --------------------------------------------------------------------------------- :   -
d05_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
d06_SOCS1R : -GGAGGGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  80
d07_SOCS1R : -GGAGGGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  80
e01_SOCS1R : -----------AGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  70
e02_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
e03_SOCS1R : --------------------------------------------------------------------------------- :   -
e04_SOCS1R : -----------AGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  70
e05_SOCS1R : -GGAGGGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  80
e06_SOCS1R : -GGAGGGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  80
e07_SOCS1R : TGGAGGGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  81
f02_SOCS1R : -------------CTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  68
f03_SOCS1R : --------AGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  73
f04_SOCS1R : -----------AGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  70
f05_SOCS1R : ----------GAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  71
f06_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
f07_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
g02_SOCS1R : --------------------------------------------------------------------------------- :   -
g05_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
g06_SOCS1R : ----------GAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  71
g07_SOCS1R : ----------GAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  71
h05_SOCS1R : ----------------------TCGCGGAGGACGGGGTTGAGGGGGATGCCCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  59
h06_SOCS1R : TGGAGGGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  81
h07_SOCS1R : -----GGGAGGAGCTCAGGTAGTCGCGGAGGACGGGGTTGAGGGGGATGCGCGCCAGGTTCTCGCGGCCCACGGTGGCCAC :  76
                        agctcaggtagtcgcggaggacggggttgagggggatgcgcgccaggttctcgcggcccacggtggccac

                    *        100        *
ref_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 117
a02_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC :  90
a03_SOCS1R : -ATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC :  35
a04_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
a05_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 107
a06_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
b01_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 106
b02_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
b03_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 106
b04_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 117
b05_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
b06_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 106
c02_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 117
c03_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 117
c04_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 106
c05_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
c06_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 106
c07_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
d01_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 107
d02_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
d03_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 106
d04_SOCS1R : -------------------TCCTGCAGCGGCCGCAC :  17
d05_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
d06_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 116
d07_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 116
e01_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 106
e02_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
e03_SOCS1R : -----------------------GCAGCGGCCGCAC :  13
e04_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 106
e05_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 116
e06_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 116
e07_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 117
f02_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 104
f03_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 109
f04_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 106
f05_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 107
f06_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
f07_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
g02_SOCS1R : ------------------------------------ :   -
g05_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
g06_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 107
g07_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 107
h05_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC :  95
h06_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 117
h07_SOCS1R : GATGCGCTGGCGGCACAGCTCCTGCAGCGGCCGCAC : 112
             gatgcgctggcggcacagctcctgcagcggccgcac
```

Figure 4.16 | Multiple alignment of SOCS box based on reverse sequencing. The predicted position of stop codon polymorphism is highlighted by the red box.

4.4 Discussion

This project began with a view to investigate the effect of polymorphisms that can be detected through clustering of ESTs. The initial focus was to examine the dataset for nonsynonymous polymorphisms (nsSNPs) and SIFT was then used to predict tolerated and deleterious nsSNPs.

The pipeline applied resulted in 5,458 nsSNPs of which 4,876 were successfully predicted to be either tolerated (2,469) or deleterious (2,407). Between them, they were assigned gene ontology terms in three broad groups - biological process, molecular function and cellular component. Subsequent analysis showed 95 ontology terms to be over-represented in the nsSNP-containing genes, including biological processes of protein folding, translational elongation, cell cycle and intracellular transport; molecular functions of antioxidant activity, protein binding, translation initiation factor activity and fatty acid binding; and to be cellular components of, among others, the cytoplasm, mitochondrion, lysosome and nucleolus.

We then posed a new question - are there SNPs that act to cause dramatic mutation(s) which lead to event(s) of adaptation?

It has long been considered that most adaptive changes in a population occur as protein changing mutations which are selected for by positive selection and subsequently become fixed within the population. Olson, however, opined that adaptive loss of function mutation may be more prevalent and spread rapidly, particularly within small populations. Loss of function mutations lead to mutated genes, he argued, but unless the genes are completely removed from genome, the mutated version may be subject to shifts in selective environments to undergo reversion. This makes gene loss a major motif in molecular evolution (Olson 1999).

In human, cases of adaptive loss of genes have been shown in immunity and pathogen resistance. Caspase-12 was driven by positive selection to maintain the stop codon polymorphism to confer resistance to severe sepsis (Xue et al. 2006). An example of stop codon polymorphism in cattle is the mutation of myostatin that leads to double-muscling in beef cattle breeds, and at the same time improve the meat quality of these cattle (Bellinge et al. 2005).

Given these known cases of beneficial loss of function mutations, affecting fitness and immunity, as well as agricultural traits, it was therefore of interest for us to identify other cases of such mutations.

SOCS1 was identified as a promising candidate following preliminary *in silico* results that gave strong evidence of a stop codon polymorphism. Discovered in 1997, SOCS1 has been shown to inhibit signalling of a wide range of cytokines including IL2, IL4, IL6 and prolactin (Larsen and Ropke 2002). To have a stop codon polymorphism in SOCS1 is therefore particularly interesting. A mutation to SOCS1 may confer an advantage in the immune response through the interleukin signalling pathway, or in lactation to increase milk yield.

Cytokines are secreted glycoproteins that are integral components of the immune system, embryonic development and haematopoiesis. They interact with cytokine receptors in order to exert their physiological effects binding to the extracellular regions of their receptors to induce receptor oligomerisation, which then acts to active Janus kinase (JAK) proteins in the cytoplasm. Specific phosphorylation of tyrosine residues in the receptor tails then takes place following JAK activation. Cytoplasmic signal transducer and activator of the transcription factors (STATs) subsequently dock to the phosphorylated receptors via its Src Homology 2 (SH2) domain and are in turn activated by phosphorylation. Dimerised STATs then translocate into the nucleus to initiate transcription. This cascade is crucial for cell development, haematopoiesis and host defence. It is also a tightly controlled and regulated cascade. Some cytokines that act via this signalling cascade include interleukins (ILs) and interferons (IFNs).

The initial analyses of our dataset involved the generation of multiple sequence alignments of genes from ESTs and the identification of stop codon polymorphism from these alignments. Of 13 genes identified to be stop codon polymorphism-containing, each were manually checked for the quality of the alignment, the presence/absence of aberrant sequences particularly at the end sequences of ESTs, the presence/absence of splice variants which could account for the variation. The trace archive of the relevant ESTs was also investigated to determine if the base-call was well supported. SOCS1 turned out to have excellent sequence alignment and showed high sequence conservation, as well as good trace records. Not only that, an additional EST which carries the same mutation was also found following a BLAST search.

Given the strength of the preliminary investigation, a decision was made to sequence around the region of the stop codon polymorphism in a panel of cattle samples to validate the presence of this mutation. However, the sequencing data did not show the presence of stop codon polymorphism that was earlier indicated in *in silico* analyses. Instead all sequences show the wild-type sequence in samples that were successfully sequenced. This led us to the conclusion that (a) this dramatic polymorphism is sufficiently rare that it doesn't exist in our sample panel, and/or (b) such dramatic mutation is unlikely to be true.

A further review of the submitted information of the ESTs to dbEST revealed that all the ESTs that helped strengthen our initial investigation were submitted by the same group of researchers. Therefore, it is possible that, a transcription error may have been reverse-transcribed and amplified by PCR prior to cloning. This would lead to an artefact, leaving multiple representations in the alignment.

Another indicator of the lack of reliability in using EST to identify SNPs came from the work by Hawken *et al.* (Hawken et al. 2004). In setting up an interactive bovine *in silico* SNP database (IBISS), they found that of 523,448 SNP detected from 48,679 multiple alignments, 285,408 of them were identified as low quality SNPs after quality screening, indicating that nearly 55% of SNPs identified in ESTs are unreliable.

Additionally, it has been reported that intolerant SNPs are rare and in the case of analyses of chicken SNPs, only 59% of the intolerant SNPs were successfully confirmed by PCR resequencing, with much of them being attributed to sequencing error, making this method still unsuitable for accurate large scale analyses (Wong et al. 2004).

This information, together with our unsuccessful validation of the stop codon polymorphism in SOCS1, despite several lines of *in silico* evidence, thus point towards the likelihood that dramatic stop codon polymorphisms are rare and are difficult to detect in ESTs. Genome-wide SNP-typing will provide insight into the frequency of such mutations in the near future.

5. Genome-Wide Signatures of Selection in Bovine Populations

5.1 Introduction

5.1.1 Cattle Domestication and its Consequences

Cattle were domesticated about ten thousand years ago, in the Neolithic age, from the aurochs (*Bos primigenius*) (Helmer et al. 2005) and with nearly one billion modern cattle raised every year, they are economically important livestock, mainly bred for beef and dairy production.

Analyses of mitochondrial DNA (mtDNA) has shown a bifurcated phylogeny between *Bos taurus* and *Bos indicus*, with a divergence time in excess of 100,000 years (Loftus et al. 1994). This led to the hypothesis that two separate domestication events took place, one which gave rise to taurine cattle and another to indicine cattle. Furthermore, previous work completed in this lab also led to the hypothesis that a third domestication centre may be implicated in the domestication of African *Bos taurus* (Troy et al. 2001). At least it is likely that the separation of European and African cattle is a primary divide in post-domestic cattle history (Beja-Pereira et al. 2006).

Domestication brought the previously wild aurochs into close human contact as well as with other domesticates, subjecting cattle to novel selective pressures including contact with pathogens that were normally present in other animals. In addition, artificial selection driven by the selective breeding of cattle for desirable traits (e.g. better milk yield, even temperament, improved meat quality) also resulted in significant genetic variation between different breeds. Moreover, adaptation to different environments and socio-cultural context also shaped genetic diversity in domesticated cattle. The identification of such selective signatures, in particular signatures of positive selection, has the potential to highlight those genomic regions that have been of most importance in the adaptation of cattle to these new selective pressures.

5.1.2 Positive Selection Within Species

The detection of positive selection between different species examines increased proportions of function altering mutations. The rate of nonsynonymous substitutions (i.e. protein-changing) is compared to the rate of synonymous substitutions (i.e. mostly silent, non-amino acid altering mutations), yielding estimates of d_N/d_S. To detect positive selection within species however, different population genetics statistical approaches are required. Single nucleotide polymorphisms (SNPs) are the most common form, of genetic variant from which signatures of selection can be detected within species.

As previously mentioned in the Introduction Chapter, Section 1.3, one signature of positive selection is a reduction in genetic diversity due to a hitchhiking effect. When a complete selective sweep takes place, the selected allele rises to fixation and with it, linked genetic variants in the region. With time, new mutations occur but they are typically rare. Some of the most commonly used statistical tests to detect regions of low diversity with an excess of rare alleles are the Tajima's *D* (Tajima 1989), and Fu and Li's *D* tests (Fu and Li 1993).

Another signal of positive selection is an excess in the frequency of derived alleles. When a non-ancestral allele arises by a new mutation, it occurs at a lower allele frequency than the ancestral allele. Following a selective sweep, however, derived alleles that are linked to advantageous alleles can rapidly rise to high frequencies and are subsequently maintained within the genome. Fay and Wu's *H* is a statistical method usually used to test for an excess of derived alleles (Fay and Wu 2000). In order to carry out this test, it is imperative that the ancestral alleles are known and this is normally inferred from the alleles found in closely related species.

A phylogeny for Bovidae is shown in Figure 5.1, showing relatedness between the cattle and other closely related species including gaur, yak, bison, buffalo and anoa. This phylogeny is reconstructed based on a neighbour-joining analysis of the *Bovinae* subfamily by MacEachern *et al.* (MacEachern, McEwan, and Goddard 2009). To this end, they sequenced 84 amplicons from 15 different genes across chromosomes 1, 2, 4 and 9. These genes are located within suspected QTL for milk and have higher than average rates of molecular evolution when comparisons were carried out between *Bos taurus*, *Bos indicus* and human.

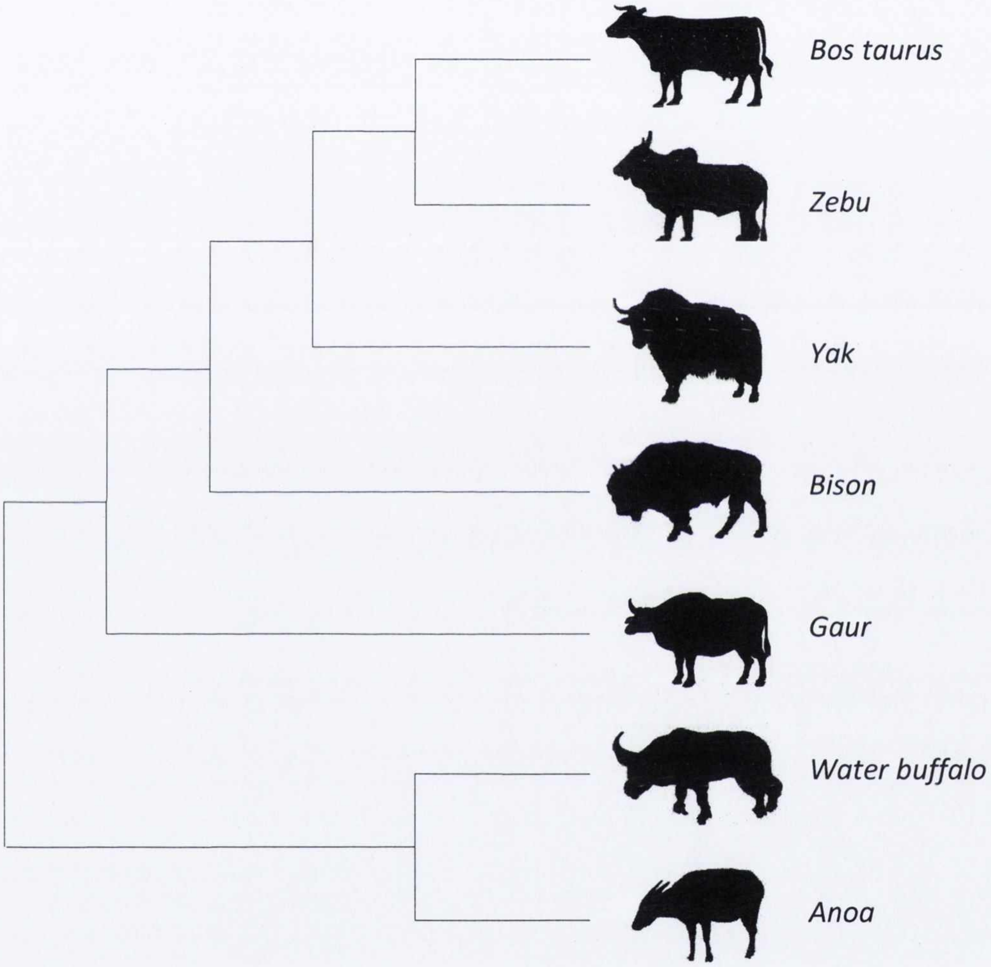


Figure 5.1 | Bovidae phylogeny to show the relationship between *Bos taurus*, *Bos indicus*, and the outgroups yak, bison, gaur, water buffalo and anoa (MacEachern, McEwan, and Goddard 2009). The outgroup species are the animals selected for SNP genotyping by the Bovine HapMap Consortium (water buffalo, anoa) and by our group (yak, bison, gaur).

When a particular species of animal is geographically separated, each population may now be subject to different selective pressures, and selection that acts on one population may not act on another. Therefore, an unusually large difference in allele frequency at a particular gene among different populations is another hallmark of selection. F_{ST} is commonly used to identify markers or regions with elevated genetic diversity between populations (Akey et al. 2002).

Commonly, when a selected allele under selection rises to fixation, it may be at a rate that prohibits substantial recombination, thus forming long-range homogeneous haplotypes. This signal can be detected using the long-range haplotype (LRH) test and the Cross Population Extended Haplotype Heterozygosity test (XP-EHH) (Sabeti et al. 2002).

5.1.3 Bovine Haplotype Map (HapMap) Consortium and 33K SNPs

The Bovine HapMap Consortium recently genotyped over 37,400 SNPs in a panel of 501 animals sampled from 19 cattle breeds (n = 497) and 2 outgroups (n = 4) to conduct a genome-wide survey of genetic variation across cattle breeds (Gibbs et al. 2009).

Among the breeds were 12 European breeds, 1 African breed, 3 Indicine breeds, and 3 hybrids. The 2 outgroups included were anoa and water buffalo. A list of cattle breeds, number of cattle genotyped, regions of origin and primary purpose(s) of the breeds is presented in Table 5.1. For further details of the technology involved in SNP genotyping, please refer to Section 5.2.1 in Materials and Methods.

Table 5.1 | Cattle populations investigated in the Bovine HapMap Project, with details of the number of animals genotyped, region of origin of breeds, and primary production use of the breeds.

	Code & Breed	N	Origin	Primary Purpose
Taurine	ANG – Angus	27	European	Beef
	BSW – Brown Swiss	24	European	Dairy
	CHL – Charolais	24	European	Beef
	GNS – Guernsey	21	European	Dairy
	HFD – Hereford	27	European	Beef
	HOL – Holstein	53	European	Dairy
	JER – Jersey	28	European	Dairy
	LMS – Limousin	42	European	Beef
	NRC – Norwegian Red	25	European	Dairy/Dual Purpose
	PMT – Piedmontese	24	European	Beef/Dual Purpose
	RGU – Red Angus	12	European	Beef
	RMG – Romagnola	24	European	Beef
	NDA – N'Dama	25	African	Multi-purpose
Indicine	BRM – Brahman	25	Indian	Beef
	GIR – Gir	24	Indian	Dairy/Multi-purpose
	NEL – Nelore	24	Indian	Beef
Hybrid	BMA – Beefmaster	24	American	Beef
	SGT – Santa Gertrudis	24	American	Beef
	SHK – Sheko	20	African	Multi-purpose
Outgroup	ANO – Anoa	2		None
	BUF – Mediterranean Buffalo	2		Dairy/Dual Purpose

One of the analyses performed by the Bovine HapMap Consortium was to examine population structure by analysing SNP genotype frequencies with the InSTRUCT software program (Figure 5.2). With the assumption of two ancestral populations (i.e. *K*=2; *Bos taurus* vs. *Bos indicus*) this analysis revealed the expected clustering of breeds - the European and African taurine cattle in one cluster with the indicine cattle in the other cluster. Hybrid animals with both taurine and indicine origins clustered between the two major groups.

At *K*=3, the African breeds grouped separately from the European cattle, in line with the theory of a probable third domestication centre of cattle in Africa (Troy et al. 2001).

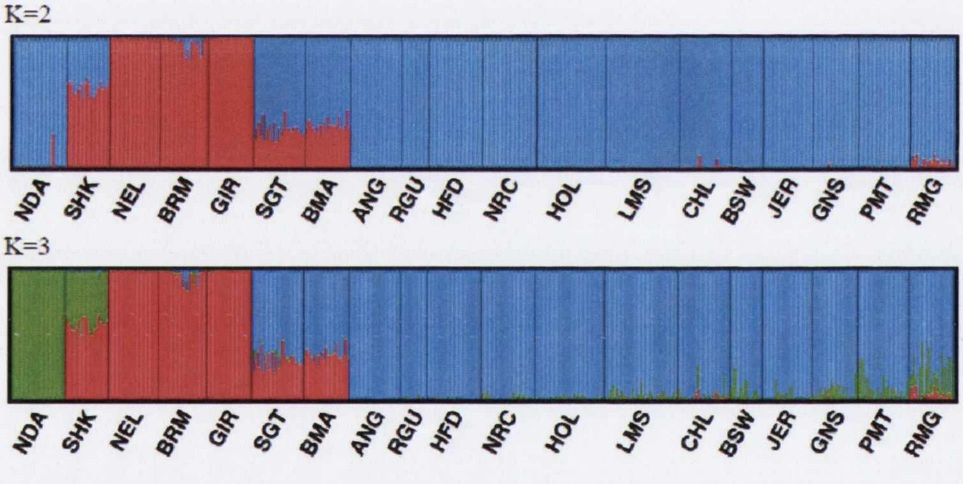


Figure 5.2 | Analysis of bovine population structure using InSTRUCT reveals the primary clustering of taurine (blue) and indicine (pink) cattle breeds at K=2. At K=3 the separation of African breeds is also apparent (Image from Bovine HapMap Consortium).

This finding is further supported by a multidimensional scaling representation of interbreed FST distances of the 33K dataset that reveals clustering of the European breeds, the separation of the African N'Dama, the clustering of the zebus and the intermediate position of the hybrid cattle breeds (Figure 5.3). This formed the basic premise for our analyses as part of the HapMap consortium, that we could study the populations of European, African and Indicine cattle as separate population groups.

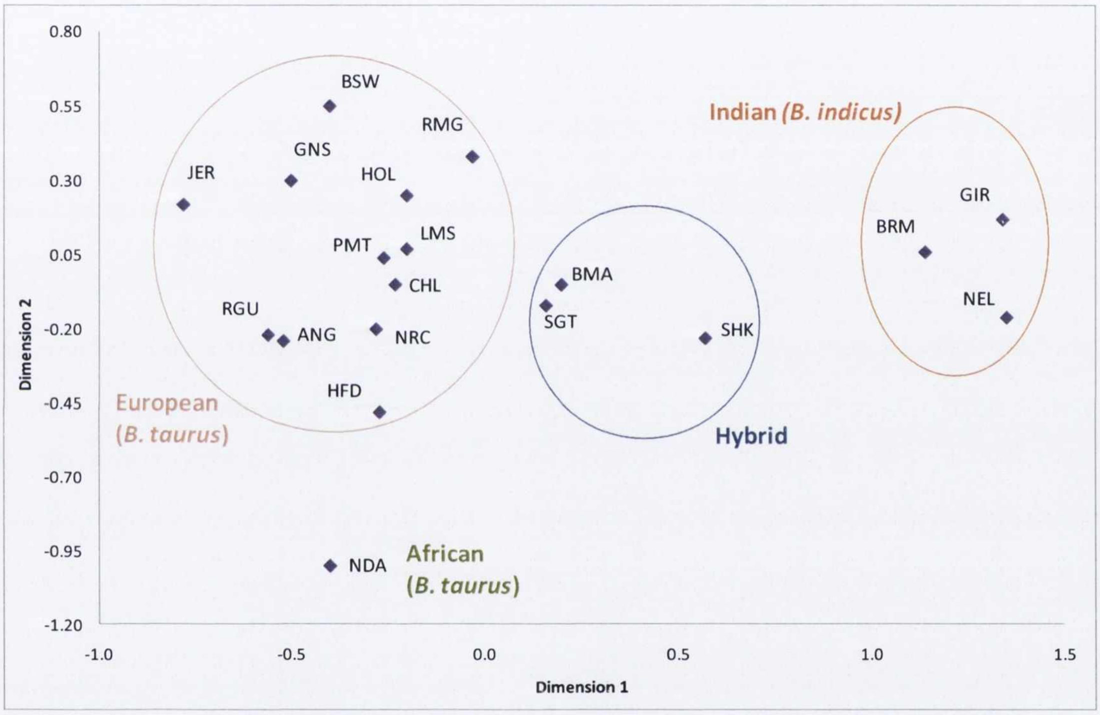


Figure 5.3 | The clustering of distinct cattle populations into European, African, Indicine and hybrid breed populations, based on multidimensional scaling of average pairwise F_{ST} distances between breeds calculated using 33K SNP genotypes.

5.1.4 Improved Bovine 50K SNPs

The availability of the first bovine genome-wide SNP genotype dataset enabled promising strategies to detect signatures of selection in cattle populations. One such strategy presented in this thesis is the investigation of SNP allele frequencies between the geographically distinct populations using F_{ST} as a measure of genetic diversity.

The initial dataset used consisted of approximately 33,000 genotyped SNPs (33K Chip) - a combination of the 10K Illumina Bovine SNP chip and the 23K ParAllele Affymetrix Genechip Bovine Genome Array. However, the dataset was not without issues. Firstly, the distribution of SNPs was not even across the chromosomes. Chromosomes 6, 14 and 25 were much more densely represented than all other chromosomes. Moreover, the locations of the SNPs were initially based on bovine genome assembly 3.0 but were subsequently mapped to the corresponding positions in bovine genome assembly 4.0. However, as the genome assembly quality improved, some of the previously erroneously placed SNPs were then no longer included in any of the chromosomes and were relegated to "position unknown", causing a loss of data-points. Additionally, the general density of the SNPs genotyped was relatively sparse, with an average SNP distribution of 1.2 SNPs per 100kb of genomic sequence.

While the genotyping and analyses efforts of the Bovine HapMap Project were underway (i.e. the 33K dataset), a bovine SNP array of higher density was being developed (Matukumalli et al. 2009). The BovineSNP50 array contains approximately 54,000 SNPs that are more evenly distributed among the chromosomes than that of 33K platform, with a mean SNP distribution of 1.98 SNPs per 100kb of sequence. Figure 5.4 illustrates the improvement of sequence mapping between bovine genome assembly 3.0 and assembly 4.0, as well as the number of SNPs per chromosome (mapped according to assembly 4.0) from both the 33K and the 50K arrays. It highlights a much more even SNP distribution for the 50K array, with improved coverage.

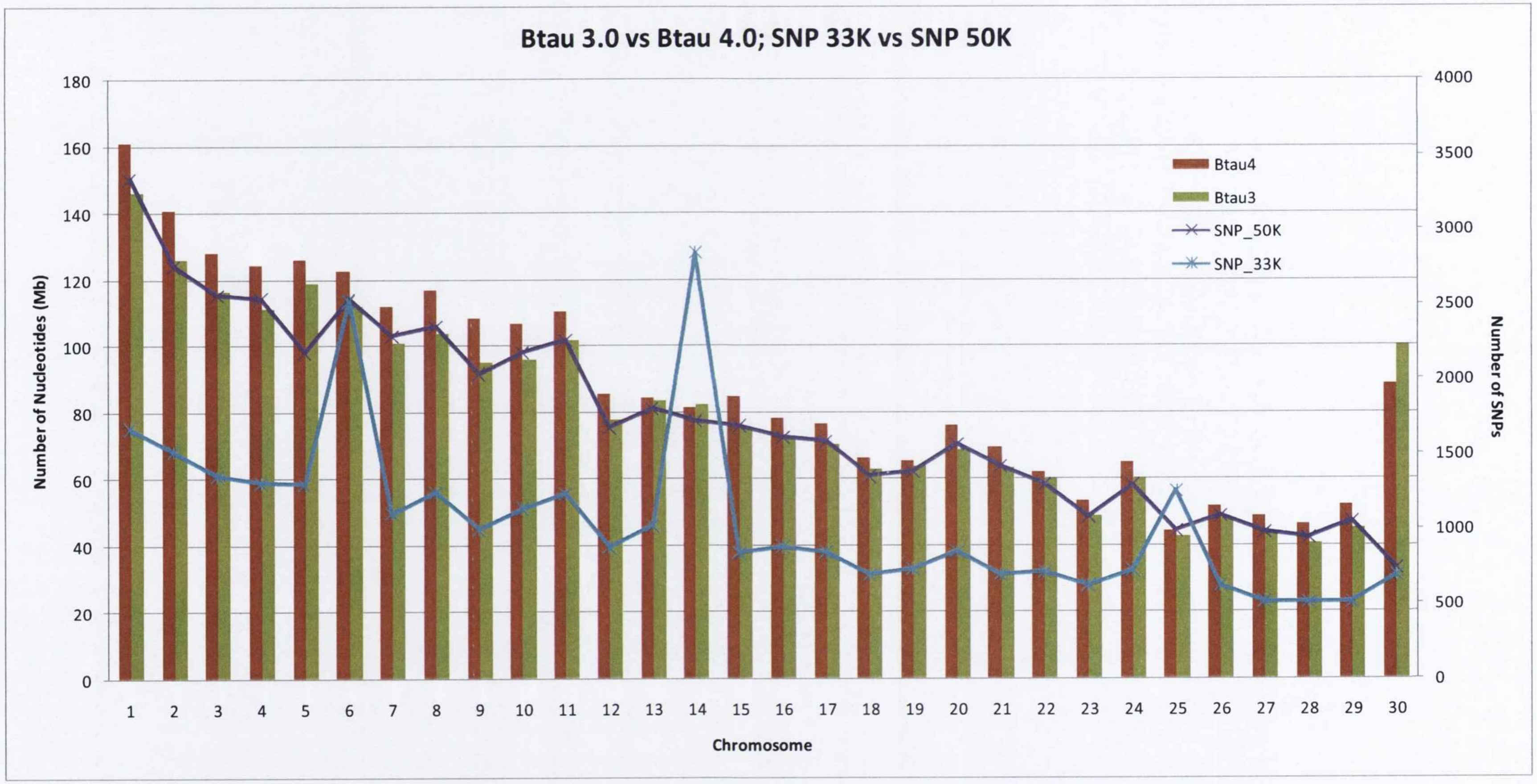


Figure 5.4 | Comparison between bovine genome assembly 3.0 and assembly 4.0, and the SNP distribution from the 33K and 50K arrays highlighting that the SNP distribution among chromosomes is more even on the 50K array. Chromosome "30" refers to Chromosome X.

5.1.5 F_{ST} as Measure of Positive Selection

F statistics were introduced by Wright as a tool to describe genetic variance and diversity within and between populations (Wright 1951). F_{ST} measures the proportion of genetic variance in a subpopulation relative to the total genetic variance, F_{IS} correlates genetic diversity between an individual relative to the subpopulation, and F_{IT} relates the allelic variance of an individual relative to the total population. The F_{ST} statistic was used in this thesis to investigate genetic diversity between each cattle subpopulation.

Under neutrality, genetic drift affects all loci in a similar manner and therefore genetic variance components should be similar among chromosomal regions. However, when positive selection occurs within one population but not the others, allele frequencies around the selected locus change and this drives a localised increase in F_{ST}. When F_{ST} is close to 1, it indicates nearly completely different genetic variance at that locus in that population relative to overall variance. A small F_{ST} value means that the allele frequencies within each population are similar, and a F_{ST} value of 0 indicates that the populations are genetically indistinguishable. Therefore by calculating F_{ST} for each genotyped marker across the genome, we can identify regions with significantly elevated or repressed diversity as outliers on the genome-wide F_{ST} distribution, and identify outliers potentially representing regions of the genome which have undergone positive selection.

Numerous studies have used the F_{ST} approach to identify natural selection, mainly in human populations. One key study was by Akey *et al.,* where they used genome-wide estimates of F_{ST} in humans to examine 26,530 loci from three populations (African-American, East Asian and European-American) and identified 174 regions that showed signatures of selection (Akey et al. 2002). Among these, 156 demonstrated unusually high levels of F_{ST} including the cystic fibrosis transmembrane conductance regulator (CFTR) gene and coagulation factor V (F5), both of which have previously been identified as genes under selection (Lindqvist et al. 1998; Slatkin and Bertorelle 2001).

With the emergence of a medium density bovine SNP genotyping platform, similar strategies can now be applied in examining cattle populations. In an early effort, MacEachern *et al.* analysed over 7,500 SNPs in beef (Angus) and dairy (Holstein) breeds, using bison, yak and banteng as the outgroups, to identify recent signatures of positive selection (MacEachern et al. 2009). However, their effort did not yield strong signatures of positive selection nor remarkable candidate genes, with the exception of fibroblast growth factor 1 (FGF1) in Angus within a QTL region, which had previously been identified for body composition and carcass yield.

More recently, Flori *et al.* genotyped 42,486 bovine SNPs using the BovineSNP50 assay in three dairy cattle breeds (Holstein, Normande and Montbéliarde) to investigate regions of the genome which may carry signatures of the intense artificial selection for traits that improve milk yield (Flori et al. 2009). Using a F_{ST}-based approach they identified 13 regions of high significance implying strong and/or recent positive selection. Some of these contain genes that have been reported to affect milk production traits (e.g. growth hormone receptor, GHR) and colouration (e.g. melanocortin 1 receptor, MC1R).

5.1.6 Intercontinental Locus Specific Branch Length (LSBL) to Identify Regions of Significant Genetic Diversity Between Populations

It has previously been postulated that there have been three major bovine domestication events; one in the Near East which gave rise to modern European cattle, one in Africa, and one in the Indo-Pakistan region that gave rise to indicine breeds (Troy et al. 2001). The divergence between these subgroups is ancient, and samples may be identified with minimal recent exchange. This is reflected in the major axes of divergence between breed samples, as shown in Figure 5.3, which was derived based on average pairwise F_{ST} distances calculated from the 33K dataset, followed by multidimensional scaling.

One indicator of loci subject to positive selection among geographically distinct population groups is maximal divergence of genomic regions between the continental groups, perhaps as a result of different selective pressures within the different domestication histories. This can be measured by Locus Specific Branch Length (LSBL), which is derived from pairwise F_{ST} distances (Shriver et al. 2004).

The LSBL approach isolates allele frequency changes geometrically, and is therefore able to quantify evolutionary rates as well as specifying the population(s) within which particular loci have undergone changes. In essence, LSBL decomposes F_{ST} here into three component parts, which are the different continental origins of the cattle breeds - African, European or Asian. This is particularly useful, because as successful an approach as F_{ST} has been, it is sensitive to changes in any one of the populations analysed (Shriver et al. 2004) and here the overall F_{ST} values may be swamped by high zebu diversity such that more subtle within-taurus differences between European and African taurine cattle may go undetected.

5.1.7 Ancestral Allele and Derived Allele Frequencies

The availability of outgroup species enabled the determination of the ancestral and the derived allele for the loci analysed in this project. This information, in turn, can be used to detect for positive selection; one signature of positive selection is an excess of derived alleles.

Derived alleles are typically expected to be present in low frequency, usually representing the minor variant. New mutations, however, may be linked to selected alleles and following a selective sweep, can rapidly rise to high frequency. Tests for selective history such as Fay and Wu's *H*, make use of ancestral allele information to determine if there is a non-neutral excess of derived alleles.

5.2 Materials and Methods

5.2.1 Bovine HapMap 33K Dataset

The Bovine HapMap dataset was provided by the Bovine HapMap Consortium. The sample collection, SNP discovery and panel design, genotyping effort, and data clean up were as discussed in the supplementary online material (SOM) that accompanied the research publication (Gibbs et al. 2009) (See Supplementary Documentation).

We also carried out several additional modifications to the data prior to analysis. There were 44 trios of dams, sires and calves included to allow verify of Mendelian inheritance as a quality control measure. Here, in order to maintain a dataset with as few unrelated individuals as possible, the alleles genotyped for the 44 calves were removed from the analysis. 10 other animals were also removed from analysis due to low genotyping success rates (<70% as opposed to all others with >90%).

We removed a further 1,032 loci from the dataset for the following reasons: loci that had more than 2 discordant trios (22); markers where at least one breed was out of Hardy-Weinberg Equilibrium (HWE) and had a genotyping error rate >=5% (393); cases where there was more than 10% missing data in >50% of *B. taurus* and >50% *B. indicus* samples (338); cases where the minor allele frequency (MAF) was <0.05 in all breeds (264); and cases where the SNP was monomorphic in all breeds (15). Taking into account all of the relevant criteria, our final dataset contained 33,849 SNPs.

5.2.2 F_{ST} Calculations and Plots

A Perl script was written to calculate F_{ST} based on the method previously published by Akey *et al.* (Akey et al. 2002).

F_{ST} is estimated from the following:

$$F_{ST} = \frac{\text{MSP} - \text{MSG}}{\text{MSP} + (n_c - 1)\text{MSG}}$$

where:

1) MSG is the observed mean square errors for loci within populations
2) MSP is the observed mean square errors between populations
3) n_c is the average sample size across samples following incorporation and correction for the variance in sample size over subpopulations.

MSG is calculated from the following:

$$\text{MSG} = \frac{1}{\sum_{i=1}^{s} n_i - 1} \sum_{i}^{s} n_i p_{Ai}(1 - p_{Ai})$$

where

1) i denotes the subpopulation (where i = 1, ..., s)
2) p_{Ai} is the frequency of SNP allele A in the ith population

MSP is calculated from the following:

$$\text{MSP} = \frac{1}{s\text{-}1} \sum_{i}^{s} n_i (p_{Ai} - \bar{p}_A)^2$$

where

1) n_i is the sample size in subpopulation i;
2) the weighted average of p_A across population is given by $\bar{p} = n_i p_{Ai} / \sum_i n_i$.

n_c is calculated from the following:

$$n_c = \frac{1}{s\text{-}1} \sum_{i=1}^{s} n_i - \frac{\Sigma_i n_i^2}{\Sigma_i n_i}$$

The range of F_{ST} value lies between 0 and 1, indicating at extremes that populations are genetically identical or completely divergent, respectively. However, in estimating F_{ST}, the number may occasionally be negative. This occurred when the mean square errors for the loci within populations (i.e. MSG values) were larger than the mean square errors between populations (i.e. MSP values). These values were reset to 0, since F_{ST} with negative values does not have a biological interpretation.

The calculation of F_{ST} generated matrices of pairwise values for all the breeds of cattle analysed in each dataset. Table 5.2 is an example of pairwise F_{ST} 19 x 19 matrix for a single SNP (BTA-107177) based on the 33K dataset.

The resulting F_{ST} matrices were then used to generate plots of average pairwise F_{ST} distributions for each chromosome. Loci that fell within the top 1% of the F_{ST} distribution were marked with red for ease of visualisation (http://www.gen.tcd.ie/llau/thesis). The data points from these plots were also linked to the Ensembl Bovine Genome Browser (http://www.ensembl.org/Bos_taurus/Info/Index) to allow the identification of genes in the regions surrounding SNPs of interest. Using the browser one can hover over each data point to display the SNP identifier, its F_{ST} value and its position on the chromosome.

Table 5.2 | 19 x 19 pairwise F_{ST} matrix for the SNP BTA-107177.

BTA107177	PMT	RGU	RMG	SHK	BSW	HOL	LMS	NDA	NEL	BMA	GIR	GNS	SGT	NRC	JER	HFD	ANG	BRM	CHL
PMT	0.0000	0.2378	0.0000	0.0075	0.0000	0.1878	0.0111	0.0000	0.0620	0.1553	0.0620	0.0524	0.0000	0.1553	0.0000	0.3517	0.2558	0.1295	0.0133
RGU	0.2378	0.0000	0.4282	0.4473	0.2804	0.0000	0.0657	0.3406	0.5352	0.0000	0.5352	0.0000	0.1441	0.0000	0.1174	0.0000	0.0000	0.6187	0.0238
RMG	0.0000	0.4282	0.0000	0.0000	0.0000	0.3094	0.1110	0.0000	0.0000	0.3110	0.0000	0.1923	0.0401	0.3110	0.0669	0.4965	0.4052	0.0278	0.1457
SHK	0.0075	0.4473	0.0000	0.0000	0.0000	0.3213	0.1251	0.0000	0.0000	0.3263	0.0000	0.2096	0.0574	0.3263	0.0828	0.5085	0.4180	0.0115	0.1644
BSW	0.0000	0.2804	0.0000	0.0000	0.0000	0.2166	0.0308	0.0000	0.0382	0.1899	0.0382	0.0810	0.0000	0.1899	0.0000	0.3865	0.2908	0.1041	0.0386
HOL	0.1878	0.0000	0.3094	0.3213	0.2166	0.0000	0.0628	0.2557	0.3701	0.0000	0.3701	0.0032	0.1211	0.0000	0.1012	0.0054	0.0000	0.4136	0.0293
LMS	0.0111	0.0657	0.1110	0.1251	0.0308	0.0628	0.0000	0.0613	0.1773	0.0231	0.1773	0.0000	0.0000	0.0231	0.0000	0.1854	0.1022	0.2280	0.0000
NDA	0.0000	0.3406	0.0000	0.0000	0.0000	0.2557	0.0613	0.0000	0.0103	0.2393	0.0103	0.1241	0.0000	0.2393	0.0171	0.4335	0.3390	0.0714	0.0785
NEL	0.0620	0.5352	0.0000	0.0000	0.0382	0.3701	0.1773	0.0103	0.0000	0.3982	0.0000	0.2822	0.1242	0.3982	0.1433	0.5668	0.4812	0.0000	0.2402
BMA	0.1553	0.0000	0.3110	0.3263	0.1899	0.0000	0.0231	0.2393	0.3982	0.0000	0.3982	0.0000	0.0800	0.0000	0.0601	0.0121	0.0000	0.4666	0.0000
GIR	0.0620	0.5352	0.0000	0.0000	0.0382	0.3701	0.1773	0.0103	0.0000	0.3982	0.0000	0.2822	0.1242	0.3982	0.1433	0.5668	0.4812	0.0000	0.2402
GNS	0.0524	0.0000	0.1923	0.2096	0.0810	0.0032	0.0000	0.1241	0.2822	0.0000	0.2822	0.0000	0.0000	0.0000	0.0000	0.1007	0.0280	0.3547	0.0000
SGT	0.0000	0.1441	0.0401	0.0574	0.0000	0.1211	0.0000	0.0000	0.1242	0.0800	0.1242	0.0000	0.0000	0.0800	0.0000	0.2694	0.1751	0.2007	0.0000
NRC	0.1533	0.0000	0.3110	0.3263	0.1899	0.0000	0.0231	0.2393	0.3982	0.0000	0.3982	0.0000	0.0800	0.0000	0.0601	0.0121	0.0000	0.4666	0.0000
JER	0.0000	0.1174	0.0669	0.0828	0.0000	0.1012	0.0000	0.0171	0.1433	0.0601	0.1433	0.0000	0.0000	0.0601	0.0000	0.2419	0.1508	0.2076	0.0000
HFD	0.3517	0.0000	0.4965	0.5085	0.3865	0.0054	0.1854	0.4335	0.5668	0.0121	0.5668	0.1007	0.2694	0.0121	0.2419	0.0000	0.0000	0.6195	0.1451
ANG	0.2558	0.0000	0.4052	0.4180	0.2908	0.0000	0.1022	0.3390	0.4812	0.0000	0.4812	0.0280	0.1751	0.0000	0.1508	0.0000	0.0000	0.5393	0.0630
BRM	0.1295	0.6187	0.0278	0.0115	0.1041	0.4136	0.2280	0.0714	0.0000	0.4666	0.0000	0.3547	0.2007	0.4666	0.2076	0.6195	0.5393	0.0000	0.3190
CHL	0.0133	0.0238	0.1457	0.1644	0.0386	0.0293	0.0000	0.0785	0.2402	0.0000	0.2402	0.0000	0.0000	0.0000	0.0000	0.1451	0.0630	0.3190	0.0000

5.2.3 Locus Specific Branch Length (LSBL) Calculation and Sliding Window Analysis

Locus Specific Branch Length (LSBL) for each locus was calculated using the corresponding values of pairwise F_{ST}. For this analysis, only non-admixed breeds were considered as these would confound the calculations. Breeds were grouped into three geographical/continental groups: European (ANG, BSW, CHL, GNS, HFD, HOL, JER, LMS, NRC, PMT, RGU, RMG), African (NDA) and Indicine/Indian (BRM, GIR, NEL).

In order to calculate LSBL, a set of average pairwise F_{ST} values was calculated for every SNP - “European-African” (EA), “European-Indicine” (EI) and “African-Indicine” (AI). For example, in the case of the “African-Indicine” F_{ST} calculation, pairwise values for BRM vs. NDA, GIR vs. NDA and NEL vs. NDA were now averaged to get a single mean “African-Indicine” value per SNP. The LSBL was then calculated for each locus in each population. Figure 5.5 is a simple schematic illustrating each branch length relative to another, and Table 5.3 lists the formulae applied to calculate the LSBLs.

Figure 5.5 | Schematic representation of LSBL branches relative to one another.

Table 5.3 | Formulae to calculate LSBL values for each population relative to two others. Each is given by the sum the pairwise F_{ST} values related to the population of interest less pairwise F_{ST} value of non-population of interest, all divided by two.

Branch Length	Formulae
ℓ_{EUR}	(European-Indicine-F_{ST} + European-African-F_{ST} - African-Indicine-F_{ST}) / 2
ℓ_{AFR}	(European-African-F_{ST} + African-Indicine-F_{ST} - European-Indicine-F_{ST}) / 2
ℓ_{IND}	(European-Indicine-F_{ST} + African-Indicine-F_{ST} - European-African-F_{ST}) / 2

Given the uneven inter-marker distance in the dataset and the objective to identify regions likely to be under selection, a sliding window analysis based on the calculated LSBL values was implemented. A mean LSBL value was obtained for every window of 5 consecutive SNPs, and a slide of 1 SNP across the chromosome was carried out. This also acted to eliminate probable false positives posed by a single SNP with high LSBL value among a region of several/numerous neighbouring SNPs with low LSBL values.

A list of all bovine genes annotated in the Ensembl database was queried using BioMart (http://www.ensembl.org/biomart/martview/) to return the accession number, gene name, gene description (if any), gene symbol (if any), and the start and end position of each gene. Genes that were located in regions with signatures of elevated or suppressed diversity were identified via the genomic position of the marker relative to the annotated genes. Sliding windows that were either completely or partially located within an annotated gene were then identified.

As the bovine gene set was poorly annotated, orthologous human genes were used to provide additional functional annotation. To do this, human genes were downloaded from GenBank (Benson et al. 2006) and a BLAST (Altschul et al. 1990) search between the Ensembl bovine genes and the GenBank human genes was carried out.

5.2.4 Bovine 50K Dataset

Concurrently, the 50K bovine SNP array was made available for genotyping a set of markers that were more evenly distributed across the chromosomes. This assay was superior to the 33K array with the median inter-marker interval reduced to 37kb (Matukumalli et al. 2009). Moreover, the genotyping effort of the 33K dataset contained a huge amount of data for European breeds but was lacking information on African and Indicine/Indian cattle. To this end, we decided to genotype a panel of cattle that represented the three geographical populations as previously discussed, using a more geographically representative set of cattle breeds.

Genotyping of the 50K dataset was carried out commercially by Aros Applied Biotechnology from Denmark (http://www.arosab.com/) using samples that were available to our lab. Three populations consisting of 111 animals from were selected for genotyping, including 69 Holstein (HOL) and Friesian (FRI) to represent European *Bos taurus*, 24 Somba (SMB) representing African *Bos taurus*, and 18 Hariana (HAR) and Tharparkar (THA) which represent Indian zebu. Another six individuals from three outgroups (i.e. 2 each) for plains bison (EBI), gaur (GAU) and yak (YAK) were also genotyped, for inferring ancestral and derived allele status.

From the full set of 54,001 markers, 122 loci were removed as they were autosomal X-linked markers, 937 were filtered due to low call rate (< 0.1) and 696 having low minor allele frequency were also removed. No markers were identified as being out of Hardy-Weinberg Equilibrium (HWE). The working dataset therefore contained 52,246 markers, but was subsequently filtered down to 43,029 markers following successful (but conservative) ancestral allele assignment (see Section 5.2.5 below for more details). These markers were then analysed using F_{ST} and LSBL calculations, following the same protocol described in the previous sections.

5.2.5 Ancestral and Derived Allele Frequencies, and Fay and Wu's *H*

Information obtained following the determination of ancestral and non-ancestral alleles can be used to test for clusters of high-frequency derived alleles, another hallmark of positive selection.

Ancestral and derived alleles were inferred from SNPs genotyped in the outgroup species – these were water buffalo and anoa for the case of 33K SNP dataset and plains bison, gaur and yak for the case of 50K SNP dataset. The inference was made based on two assumptions, that (a) the mutation at a particular locus happened only once, and (b) the mutation occurred following the speciation of cattle and the outgroups. However, in view of the low call rate in the two buffalo species, probably a consequence of their greater phylogenetic distance, we decided not to use the assignments of ancestry in the 33K data set.

The 50K dataset was assigned ancestry. In order to maintain the most conservative dataset of assigned ancestral and derived alleles, a successful inference was considered only where there was at least one (out of two) successful homozygote genotype calls within each species pair of outgroups. Any loci with outgroup heterozygote calls were eliminated from the dataset. As we have only a pair of animals per outgroup species, in order to minimise false positive arising from sequencing error, loci with disagreements between animals from the same species were also excluded from subsequent analysis. Therefore, our dataset is one of conservative nature. Once the ancestral allele was determined, the alternative allele was assigned to be the derived allele. Both the ancestral and the derived allele frequencies were then calculated.

One test for excess of derived allele is Fay & Wu's *H* (Fay and Wu 2000). The *H* statistics can be expressed in terms of ancestral and derived allele frequencies, as per the method for statistical analysis carried out by MacEachern *et al.* (MacEachern et al. 2009).

Fay and Wu's *H* is given by: $H = \theta_{\pi} - \theta_{H}$

where:

1) θ_{π} is the measure of heterozygosity

2) θ_{H} is the measure of homozygosity

They each can be given by: $\theta_{\pi} = \frac{\Sigma 2p\,(1-p)}{N_L}$ $\theta_{H} = \frac{\Sigma 2p^2}{N_L}$

where:

1) p refers to the derived allele frequency

2) 1- p refers to the ancestral allele frequency

3) N_L refers to the number of loci

5.2.6 Composite Empirical Calculation

In an attempt to evaluate both the LSBL and Fay and Wu's *H* together, a composite empirical method was applied. In order to do so, results from two measures were set to comparable scales, through ranking by percentile. As Fay and Wu's *H* is based on differences, with the most negative values indicative as strong signals of selections, data transformation was necessary.

First, the negative values were converted to positive ones, and vice versa. Subsequently, a fixed small number would be added to the entire distribution so all values for this new distribution would be positive, with the highest values set to be ranked highly as well. The new dataset was then ranked by percentile. Once both measures were ranked, at each position where a marker was present, the product of the corresponding values was taken, giving a final composite empirical calculation. Sliding windows which fell within the top percentile of the distribution were identified, and the genes within these windows were investigated as candidates of positive selection.

5.3 Results

5.3.1 Bovine 33K SNP - LSBL Analysis

Our data were provided by the Bovine HapMap Consortium. While the original dataset contained over 37,400 markers, following a series of quality control, a total of 33,849 markers remained in the dataset for analyses. A further 2,003 markers were subsequently discarded from the dataset, as these markers cannot be mapped to the current build of the bovine genome, thus 31,846 markers remained for analysis.

Pairwise F_{ST} values were calculated for each breed of cattle compared to each other breed. These pairwise F_{ST} values were subsequently used in the calculation of locus specific branch length (LSBL). The calculation of the LSBL was carried out on a sliding window basis, each window containing 5 SNPs and each slide is by 1 SNP.

Tables 5.4, 5.5 and 5.6 tabulate the top 20 contiguous windows within the African, European and Indian cattle populations respectively. In building the contiguous windows, the SNP windows from the top 1% distribution were initially sorted by the maximum LSBL value per window. Then, any adjacent windows to those with the highest LSBL values were clustered together to form contiguous windows. Both the number of contiguous windows clustered per result and the highest LSBL value per contiguous window are included in the result tables. The cut off LSBL values for these populations were 0.2763 for the top 1% for European cattle, 0.3003 for African cattle and 0.5208 for the Indian cattle (Supplementary Table 5.1).

The 33K SNP dataset was not used extensively to investigate for derived allele frequency nor to attempt a calculation of Fay and Wu's *H* statistics. Just over 11,000 SNPs were confidently assigned their ancestral allele and among these, only 10,427 loci could be successfully mapped to the chromosomes.

Table 5.4 | 33K sliding window analysis for the African cattle population. The number of contiguous windows that were clustered together is indicated and the top LSBL value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Win_Start	Win_End	No of Win	Afr_Top_LSBL	Genes
14	26175603	26216303	9	0.6386	CHD7
1	115442246	115830463	7	0.5878	GPR149, DHX36, SGEF
14	11099144	11173779	12	0.5233	
17	810185	1590263	6	0.5127	
9	45052194	46502866	5	0.4817	RTN4IP1, AIM1, QRSL1, ATG5, RPS3A, PRDM1
6	88335937	88419757	19	0.4787	CSN2, STATH, CASA2
1	109643161	110567887	10	0.4567	IQCJ, MFSD1
13	70143146	70801373	4	0.4452	TOP1, PLCG1, ZHX3, LPIN3, EMILIN3, CHD6
18	14857895	14984723	6	0.4362	ITFG1
9	65307268	65977474	3	0.4349	SYNCRIP, SNX14, NT5E, RPL22L1
25	6890893	6932793	6	0.4294	
22	32556552	32753508	5	0.4281	MITF
13	5391807	5593520	4	0.4252	BTBD3
14	5535813	5568995	4	0.4236	
28	3760295	4231001	5	0.4152	KIAA1383, C1orf57, PCNXL2
27	17241713	17432280	3	0.4150	SORBS2
20	23810982	24231065	5	0.4140	MIER3, C5orf35, MAP3K1
11	39857711	40187418	3	0.4117	SNRPEL1, EFEMP1
14	53068996	53150870	2	0.4039	TRHR
2	23269908	24111811	3	0.4036	OLA1, SP3, CDCA7

Table 5.5 | 33K sliding window analysis for the European cattle population. The number of contiguous windows that were clustered together is indicated and the top LSBL value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Win_Start	Win_End	No of Win	Eur_Top_LSBL	Genes
14	10878951	10953660	6	0.5464	
25	43217584	43490436	5	0.4775	PSMG3, TMEME184A, MAFK, INTS1, MICALL2
25	2513906	2611540	6	0.4741	ATP6V0C, AMDHD2
3	78870020	79585346	10	0.4568	
6	37792671	37876749	6	0.4529	
13	201362	781640	3	0.4391	OR4C11, KLF17, TXNDC13
11	13907535	14161069	5	0.4285	PAIP2B, NAGK, TEX261, ANKRD53, ATP6V1B1, VAX2
22	19684073	20218940	4	0.4285	GRM7
14	10651019	10685489	5	0.4220	
24	50125408	50745615	4	0.4184	KIAA0427, SMAD7, DYM
6	52785039	53312444	13	0.4132	
15	38740265	39760998	6	0.4109	TEAD1, PARVA, MICALCL,MICAL2, DKK3, USP47
19	40538774	41490369	5	0.4105	MLLT6, PCGF2, PSMB3, PIP4K2B, CCDC49, RPL23, LASP1, FBXO47, PLXDC1, ARL5B, CACNB1, RPL19, STAC2, TFCP2L1, FBXL20, CRKRS, MED1, NEUROD2, PPP1R1B, STARD3, TCAP, PNMT, PERLD1, ERBB2, C17orf37, GRB7, IKZF3, ZPBP2
19	51616876	52285528	4	0.4042	HEXDC, FLJ22222, UTS2R, DUS1L, FLJ35767, CSNK1D, SLC16A3, FASN
14	11122530	25493914	2	0.3974	
11	39253966	39557567	4	0.3947	RTN4, FLJ31438, RPS27A, MTIF2, LOC244405, CCDC88A
14	25002378	25038628	7	0.3842	TOX
25	4194945	4257353	15	0.3835	DNAJA3, NMRAL1, HMOX2, C16orf5
19	7324450	7702072	4	0.3801	MSI2
8	14721282	15497896	4	0.3773	

Table 5.6 | 33K sliding window analysis for the Indicine cattle population. The number of contiguous windows that were clustered together is indicated and the top LSBL value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Win_Start	Win_End	No of Win	Ind_Top_LSBL	Genes
12	25363434	26132030	12	0.8635	NBEA, MAB21L1, VPS24
X	40351793	44511776	20	0.8558	POU3F4, BRWD3, NSBP1, SH3BGRL, HTR2C, PLS3
12	28881869	29188226	4	0.8330	RXFP2
2	74617145	75747118	4	0.8230	RALB, INHBB, EPB41L5, DBI, TMEM37, SCTR, MGC33657, TMEM177, PTPN4
14	9876276	9890085	3	0.7983	DDEF1
X	47719724	50263608	8	0.7943	RPS4X, ERCC6L, ACRC, CXCR3, TAF1, OGT, RAC1, ITGB1BP2, NONO, ZMYM3, GJB1, HNRPF, NLGN3, MED12, IL2RG, LOC158830, FOXO4, SNX12, SLC7A3, TEX11, DLG3, GDPD2, KIF4A, PDZD11, ARR3, DGAT2L3, DGAT2L6
19	47652056	48244114	5	0.7828	KIAA1267, CDC27, MYL4, NUFIP2, ITGB3, C17orf57
7	50991839	51659786	4	0.7539	IK, CD14, HARS, HARS2, ZMAT2, SIL1, CTNNA1, PCDHAC2, ZMAT2
X	25801187	28159624	5	0.7507	PCDH11X, SIAH1
X	21572594	23508945	5	0.7411	GABRQ, FATE1, CNGA2, MAGEA2B, GABRE, MAGEA10, GABRA3, CETN2, NSDHL, ZNF185, ZNF275, UCHL5IP, BGN, DUSP9, PNCK, SLC6A8, BCAP31, ABCD1, PLXNB3, SRPK3, IDH3G, SSR4, PDZD4, L1CAM, HCFC1, IRAK1, MECP2
17	68412425	68672943	3	0.7377	PIWIL3, SGSM1, bA9F11.1
2	133894182	134519216	5	0.7246	EPHB2, LUZP1, AOF2, LOC646262, ZNF436, HNRNPR
24	16702404	16735480	3	0.7175	
14	10632341	10647830	4	0.7147	
6	34326491	34391775	4	0.6860	
2	64368952	64793167	4	0.6845	DARS, MCM6, LCT, UBXD2, R3HDM1
19	45395513	46157261	3	0.6841	MGC34829, C17orf53, ASB16, TMUB2, ATXN7L3, UBTF, SLC4A1, RUNDC3A, SLC25A39, GRN, ITGA2B, FZD2, FLJ35848, EFTUD2, CCDC43, DBF4B, ADAM11, GJA7, HIGD1B, CCDC103, GFAP, LOC146909, C1QL1, DCAKD
24	44978575	45399708	3	0.6816	C18orf1, C18orf19, RNMT, MC5R, MC2R
20	71369875	71761741	4	0.6813	KIAA0947, ADAMTS16
16	22962769	23425355	4	0.6803	KIAA1822L, TAF1A, MIA3, C1orf80, C1orf58, DISP1

In categorising the top 1% LSBL windows by Gene Ontology in the three cattle populations, we detected only a very small number of GO terms that have significant *P*-values following multiple testing correction using Benjamini and Bonferroni test. In the Africa cattle population, over-represented categories of molecular function were protein binding and structural constituent of eye lens. In the European cattle population, keratin filament as a term under cellular component was noted to be significant. In the Indicine cattle population, biological processes of cell adhesion and multicellular organismal development were also over-represented. (See Supplemental Tables 5.2, 5.3 and 5.4 - significant terms are shaded in blue, while terms remaining significant after multiple correction are shaded in pink.)

Examination of the GO analyses for significant terms without multiple correction shows that only 5 categories of biological processes were significant aross two populations (i.e. response to other organism, integrin-mediated signalling pathway, salivary gland development, visual perception and muscle thick filament assembly) and none across all three populations. Two molecular function terms were enriched across all three populations - protein binding and SH3/SH2 adaptor activity - and nine molecular function terms were shared by two cattle populations, including chromatin binding, DNA-directed DNA polymerase activity, epidermal growth factor receptor activity and structural molecule activity.

5.3.2 Bovine 50K SNP - LSBL, Fay and Wu's *H*, and Composite Analyses

The full dataset from our own 50K genotyping experiment yielded a total of 52,246 informative SNPs for F_{ST} and LSBL analyses. With this dataset, we also conducted Fay and Wu's *H* statistical tests, as well as performing a simple composite analysis to investigate the concordance of results between the two methods.

We have included outgroup species that are phylogenetically closer to the cattle than either buffalo and anoa, which were genotyped by the Bovine HapMap Consortium, in our panel of animals used. This proximity was reflected in call rates of 97.3%, 97.2%

and 96.9% for Plains bison, gaur and yak, respectively. This approach was shown to be promising, as we could confidently assign ancestry to the alleles based on stringent criteria and still have an approximately 82% successful call rate across all three outgroup species. 43,029 of the loci were homozygous, with the genotype consistently the same allele across all three outgroup species.

In order for us to compare the results for both LSBL and Fay & Wu's *H* analyses, it was decided that these 43,029 loci would comprise the final dataset used in these analyses. The protocol for calculating LSBL remained the same as per the 33K data analysis, with sliding window performed across 5 SNP windows. Annotated genes were identified in significant windows.

To identify outlying genes of interest, the top 1% windows were sorted by the highest ranked value in the distribution. Other windows that were adjacent to these highly ranked windows were then clustered together to form contiguous window clusters. Tables 5.7 to 5.15 summarise the top 15 ranked contiguous windows from analyses of the 50K dataset. Tables 5.7 to 5.9 relate to the African population, in terms of LSBL, Fay & Wu's *H* and composite empirical analyses, respectively. Similarly, Tables 5.10 to 5.12 refer to European population whereas Tables 5.13 to 5.15 refer to Indicine/Indian populations. (For full result sets, please see Supplementary Tables 5.5, 5.6 and 5.7.)

GO term analyses, as per described in earlier chapters, were also carried out on the 50K dataset for each cattle population. Four separate analyses per population were performed - one for each of the top hits of LSBL, Fay & Wu's *H* and composite analyses, and one encompassing all the top hits from the three analyses in one single ALL list. Similar to the 33K GO analyses, a relatively large number of GO categories were identified as over-represented categories for the positive selection studies. However, after multiple correction tests was carried out, the numbers quickly whittled down to one or two in some dataset, and none in most of them. (See Supplementary Tables 5.8, 5.9 and 5.10.)

Figure 5.7 | 50K sliding window LSBL analysis for the African cattle population. The number of contiguous windows that were clustered together is indicated and the rank of LSBL value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Start	End	No of Win	Rank Afr_LSBL	Genes
3	57339352	57315260	4	0.9993	GBP3, GBP5, GBP6
17	1050809	1177912	5	0.9986	
11	97719225	97830382	4	0.9985	DENND1A, CRB2
3	124441054	124554975	4	0.9985	COL6A3, MLPH, PRLH, LRRFIP1, RAB17
7	60703866	60855463	5	0.9982	PPRAGC1B, PDE6A, LOC728287, SLC26A2, CSF1R, PDGFRB
8	76562779	76825888	6	0.9980	GNRH1, CDCA2, KCTD9, DOCK5, CDCA2
1	152334371	152817113	4	0.9978	HLCS, DSCR6, TTC3, PIGP, DSCR3
14	10052443	10322348	5	0.9976	DDEF1, FAM49B, MLZE
29	42112879	42226330	3	0.9973	FTH1, BEST1, RAB3IL1, INCENP
10	45698672	45799476	4	0.9973	PIF1, PLEKHQ1, OAZ2, ZNF609, TRIP4, BANF1
7	17583793	17692784	4	0.9973	JMJD2B, PTPRS
28	3191635	3553189	7	0.9971	SIPA1L2, KIAA1804
1	1043153	1121300	4	0.9970	SON, DONSON, CRYZL1, GART, C21orf55, TMEM50B
9	45373250	45590254	4	0.9968	AIM1, RTN4IP1, ATG5, QSRL1, PRDM1
8	86397567	86544298	5	0.9958	PTCH1

Figure 5.8 | 50K sliding window Fay & Wu's *H* analysis for the African cattle population. The number of contiguous windows that were clustered together is indicated and the rank of *H* value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Start	End	No of Win	Rank Afr_H	Genes
6	52811751	53104687	7	0.9980	
13	75535776	75599043	3	0.9990	ZNF335, SLC12A5, NCOA5, MMP9, CD40, PLTP
11	29051434	29134031	2	0.9985	SRBD1
1	64853867	64904331	3	0.9983	C3orf30, UPK1B
17	45806387	45913105	3	0.9980	TDO2, CTSO
23	4513388	4585740	3	0.9980	PRIM2
10	45359878	45799476	9	0.9973	NDUFB8, PTGDR, NID2, PIF1, PLEKHQ1, OAZ2, ZNF609, TRIP4, BANF1
3	36543962	36754026	4	0.9973	SORT1, ATXN7L2, GNAI3, GNAT2, AMPD2, AMIGO1, SYPL2, CYB561D1, CELSR2, PSRC1, MYBPHL, PSMA5
3	64805904	65085635	5	0.9972	
13	67109440	67274436	3	0.9970	BLCAP, NNAT, CYNNBL1
5	33136024	33277669	4	0.9968	KIAA1602, FMNL3, BCDIN3D, TEGT, SPATS2, KCNH3
6	38638963	39159588	8	0.9965	
27	29511612	29631656	4	0.9963	NRG1
1	159495208	159622630	4	0.9960	
28	3327703	3420102	3	0.9960	SIPA1L2

Figure 5.9 | 50K sliding window composite analysis for the African cattle population. The number of contiguous windows that were clustered together is indicated and the rank of composite empirical value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Start	End	No of Win	Rank Afr_Comp	Genes
3	57263996	58584905	11	0.9728	GBP3, GBP4, GBP5, GBP6, GBP7, CCBL2
28	3252259	3553189	6	0.9874	SIPA1L2, KIAA1804
25	2462997	2613040	4	0.9863	AMDHD2, TBC1D24, C16orf59, ATP6V0C, ABCA3, PDPK1
11	40409764	40729584	5	0.9862	CCDC85A
5	33136024	33277669	4	0.9861	KIAA1602, FMNL3, BCDIN3D, TEGT, FMNL3, C12orf25, SPATS2
10	45640747	46001271	6	0.9851	PIF1, PLEKHQ1, OAZ2, ZNF609, TRIP4, BANF1, CSNK1G1
1	152334371	152817113	4	0.9838	HLCS, DSCR6, TTC3, PIGP, DSCR3
1	158831291	159077466	7	0.9831	SATB1
8	100324743	100395004	3	0.9831	FCMD, TMEM38D
2	22308433	22381259	3	0.9821	ATP5G3, ATF2
22	55819706	55912637	4	0.9808	ATP2B2
11	83688145	83979735	5	0.9803	FAM49A
3	36603946	36797994	3	0.9801	SORT1, ATXN7L2, GPR61, GNAI3, AMIGO1, SYPL2, CYB561D1, CELSR2, PSRC1, MYBPHL, PSMA5
29	42079321	42226330	4	0.9755	FTH1, BEST1, FADS3, RAB3IL1, INCENP
14	59409760	59556106	3	0.9725	FZD6, SLC25A32, BAALC, CTHRC1, ATP6V1C1

Figure 5.10 | 50K sliding window LSBL analysis for the European cattle population. The number of contiguous windows that were clustered together is indicated and the rank of LSBL value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Start	End	No of Win	Rank Eur_LSBL	Genes
4	53947199	54139539	6	0.9988	CAV1, CAV2, TES
4	79623055	79701138	3	0.9983	OGDH, NUDCD3, DDX56, TMED4, CAMK2B
8	86000953	86297121	4	0.9980	FANCC, LOC727969
8	111309165	112094471	10	0.9979	TRIM32, ASTN2
1	85182073	85746257	7	0.9979	YEATS2, MAP6D1, PARL, ABCC5, RPS6, KLHL24, MCF2l2, B3GNT5
22	53431058	53691278	5	0.9978	CSPG5, TMEM10, PTPN23, KLHL18, SCAP, KIF9
2	123298594	123425250	4	0.9978	PRMT1, B3GNT7, NMUR1, ARMC9
1	1479888	1616279	6	0.9977	IFNAR2, OLIG2
8	56465727	56742648	6	0.9975	GNAQ, CEP78, PSAT1
14	22563085	22967674	7	0.9973	PLAG1, TMEM68, TGS1, RDHE2, MOS, LYN, CHCHD7
29	44170538	44392573	5	0.9972	MACROD1, KCNK4, ESRRA, PRDX5, PLCB3, BAD, FKBP2, TRPT1, NUDT22, CCDC88B
26	19429867	19551065	6	0.9970	
29	46218318	46367545	5	0.9968	CNIH2, YIF1A, SLC29A2, NPAS4, MRPL11, RIN1, BRMS1, B3GNT1
3	9752145	10383301	7	0.9967	SLAMF7, LY9, COPA, NCSTN, VANGL2, SLAMF1, SLAMF6, CD84, NHLH1, PEX19
1	81856895	82245378	6	0.9967	TMSB4X, ST6GAL1, RTP1, ADIPOQ, RFC4

Figure 5.11 | 50K sliding window Fay & Wu's *H* analysis for the European cattle population. The number of contiguous windows that were clustered together is indicated and the rank of *H* value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Start	End	No of Win	Rank Eur_H	Genes
X	42851846	43597178	5	0.9986	CHM, POU3F4, BRWD3, SH3BGRL, NSBP1
21	28467237	28667219	5	0.9988	TJP1, TARSL2, TM2D3
18	13260824	13495660	4	0.9988	CDH15, LOC146429, LOC197322, ANKRD11, CDK10, SPG7, RPL13, CPNE7, C16orf7, DPEP1, SPATA2L, ZNF276, FANCA
4	53972253	54098586	4	0.9987	CAV1, CAV2
14	22563085	22934037	6	0.9980	PLAG1, TMEM68, TGS1, RDHE2, MOS, LYN, CHCHD7
16	21252383	21329255	5	0.9980	TGFB2
12	34005356	34516267	5	0.9978	ATP8A2, FAM123A, SPATA13
7	49309421	50106890	5	0.9972	GFRA3, CDC25C, MATR3, CDC23, REEP2, EGR1, ETF1, HSPA9, CTNNA1, FAM53C, PAIP2, SLC23A1, CXXC5, ECSM2
10	75745784	75872732	5	0.9972	PRKCH
8	9456	26638	5	0.9972	HIATL1
9	105265310	105408677	5	0.9972	RPS6KA2
7	95426992	95640021	5	0.9966	ANKRD32, MCTP1
16	56196285	56332127	4	0.9965	FAM5B
11	29001108	29134031	3	0.9963	SRBD1
8	111482441	111689122	7	0.9963	TRIM32, ASTN2

Figure 5.12 | 50K sliding window composite analysis for the European cattle population. The number of contiguous windows that were clustered together is indicated and the rank of composite empirical value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Start	End	No of Win	Rank Eur_Comp	Genes
4	53947199	54249891	7	0.9841	CAV1, CAV2, TES
14	22563085	22967674	7	0.9933	TMEM68, TGS1, PLAG1, RDHE2, MOS, LYN, CHCHD7
8	86000953	86297121	4	0.9905	FANCC, LOC727969
20	21295487	21758189	6	0.9846	RAB3C
8	108801252	108856282	3	0.9827	COL27A1, AKNA, DFNB31
12	34005356	34516267	4	0.9825	ATP8A2, FAM123A, SPATA13
9	25735289	25805685	3	0.9821	
8	111259654	112126740	13	0.9818	TRIM32, ASTN2
24	35833878	36033656	5	0.9811	GREB1, ESCO1, PLP2
16	56128931	56393931	6	0.9810	FAM5B
29	44170538	44392573	5	0.9805	MACROD1, KCNK4, ESRRA, PRDX5, PLCB3, BAD, FKBP2, TRPT1, NUDT22, CCDC88B
28	24379487	24719965	6	0.9796	CXXC6, DDX21, KIAA1279, STOX1, DDX50, CCAR1, SUPV3L1
8	56465727	56742648	6	0.9778	GNAQ, CEP78, PSAT1
7	70491678	70636975	5	0.9757	RNF145, UBLCP1, IL12B
1	1523651	1616279	5	0.9752	OLIG2

Figure 5.13 | 50K sliding window LSBL analysis for the Indicine cattle population. The number of contiguous windows that were clustered together is indicated and the rank of LSBL value per cluster is given. Blank “Genes” column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Start	End	No of Win	Rank Ind_LSBL	Genes
X	42697584	45039107	8	0.9978	CHM, POU3F4, FGF16, SH3BGRL, ATRX, BRWD3, C10orf84, LOC644756, NSBP1
7	49309421	50106890	5	0.9984	DNAJC18, TMEM173, UBE2D2, MATR3, PAIP2, SLC23A1, LOC202051, CTNNA1, ECSM2, GFRA3, CXXC5
5	49933341	50069068	3	0.9983	DYRK2, CAND1
20	14275029	14619831	4	0.9980	ERBB2IP, SFRS12, NLN, RPS6, SGTB
13	5366935	5480552	4	0.9980	BTBD3
20	14967717	15183218	5	0.9976	PPWD1, TRIM23, ADAMTS6
5	117215497	117300831	4	0.9975	CSNK1E, KDELR3, DDX17, DMC1, CBY1, TOMM2, JOSD1, GTPBP1, LOC651105
8	9597	34574	5	0.9974	HIATL1
8	87368037	87494137	5	0.9974	ZNF367, HABP4, CDC14B
X	40437855	41471339	4	0.9973	PLS3, HTR2C
11	28909667	29134031	5	0.9973	SRBD1
13	40531620	40659125	5	0.9968	C20orf19, XRN2
13	67109440	67250167	4	0.9968	BLCAP, NNAT, CTNNBL1
X	49279034	50192453	6	0.9967	RAC1, FOXO4, TAF1, SLC7A3, IL2RG, NLGN3, ITGB1BP2, KIF4A, TEX11, GJB1, MED12, DGAT2L6, DGAT2L3
6	76626037	76831000	3	0.9967	

Figure 5.14 | 50K sliding window Fay & Wu's *H* analysis for the Indicine cattle population. The number of contiguous windows that were clustered together is indicated and the rank of *H* value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Start	End	No of Win	Rank Ind_H	Genes
8	7131386	7497689	5	0.9950	ADAM29, GLRA3, FDFT1, CTSB, NEIL2, GATA4
5	26809060	27163140	4	0.9985	CCDC41, PLXNC1, TMCC3
18	14671545	15052080	6	0.9983	GPT2, DNAJA2, ITFG1, NETO2, PHKB
8	74105201	74208708	4	0.9980	LOXL2, SLC25A37
14	27809163	27947545	5	0.9980	GGH, TTPA
1	4975	14038	2	0.9980	
1	82843582	83017465	5	0.9978	DGKG, ETV5
10	85213922	85402420	4	0.9978	MAP3K9, PCNX
7	62174884	62525452	5	0.9976	CCDC69, SLC36A3, GM2A, FAT2, SLC36A1, SLC36A2, SPARC
14	13812895	13939656	4	0.9965	
17	45698876	45913105	5	0.9964	TDO2, CTSO
2	131782872	132265351	6	0.9962	LDLRAP1, TMEM50A, RHD, SYF2, TMEM57, C1orf63, MAN1C1
4	64295179	64567621	8	0.9961	TBX20, DPY19L1, DPY19L2, NPSR1
6	52573130	53162427	13	0.9961	
22	57984634	58127890	5	0.9960	TMEM40, CAND2, RAF1, MKRN2

Figure 5.15 | 50K sliding window composite analysis for the Indicine cattle population. The number of contiguous windows that were clustered together is indicated and the rank of composite empirical value per cluster is given. Blank "Genes" column indicates that there are currently no genes annotated in that region of the bovine genome.

Chr	Start	End	No of Win	Rank Ind_Comp	Genes
X	48727556	49996809	9	0.9593	OGT, RAC1, FOXO4, TAF1, SLC7A3, IL2RG, NLGN3, ITGB1BP2, KIF4A, TEX11, GJB1, MED12, DGAT2L3, DGAT2L6
2	73773944	73908380	5	0.9847	EN1
24	57918055	57966243	2	0.9801	
4	10117218	10228074	3	0.9781	LOC442710, DKFZP564O0523, PEX1
18	14671545	14882090	3	0.9780	GPT2, DNAJA2, ITFG1, NETO2
5	45035031	45346212	4	0.9739	KIF21A
8	1078439	1410670	5	0.9690	SH3RF1, NEK1, CLCN3, C4orf27, NEK1, GRPEL2
19	27837528	27914431	5	0.9687	TNFSF12, SENP3, CD68, MPDU1, TP53, SAT2, SHBG, SOX15, FXR2, ATP1B2, DNAH2, WDR79
24	32226383	32395445	5	0.9680	
12	24343133	24815075	7	0.9670	SMAD9, ALG5, EXOSC8, RFXAP, CCNA1, SPG20
7	88936675	89083103	4	0.9669	
2	92309639	92419325	4	0.9608	SATB2
X	45855932	46565224	5	0.9590	KIAA2022, UPRT, ABCB7, RNF12, SCL16A2
3	111821915	111914286	3	0.9585	HIVEP3, EDN2, FOXO3
14	27886654	27925450	2	0.9580	GGH, TTPA

5.4 Discussion

The detection of signatures of selection using genome-wide datasets has experienced an explosion of activity within the last five years. This field has been enabled by new SNP genotyping technologies and is now gearing toward the challenge of large numbers of genomes being re-sequenced in populations of individuals from a given species. However, the inference of Darwinian selection is not a trivial task and involves multiple approaches, often with conflicting and at the least, non-overlapping results.

Nielsen *et al.* in a recent review summarised the concordance between six human genome-wide searches for selected loci, each using different approaches, and found this to be poor (Nielsen et al. 2007). The different individual studies listed from 59 to 802 selected genes, while those confirmed between approaches tended to range in single or, at most, low double figures. It has proved very difficult to isolate true findings from noise within these and subsequent large lists of candidate loci.

One reason cited for a lack of concordance is that examination of genome-wide SNP data is carried out by a range of techniques that use different features of the data as a basis (Nielsen et al. 2007). These include extended haplotype homozygosity (EHH) analysis, examination of divergence between populations, detection of the frequency of derived alleles, and a range of methods summarising aspects of the site frequency spectrum. The medium SNP density of currently available bovine data presented here precludes a reasonable EHH analysis – the same analysis approach in humans only yielded interesting results once at least 300K SNP data were available (Voight et al. 2006).

Additionally, a major issue with both the 33K and the 50K cattle datasets is that of ascertainment bias. During principal component analysis (PCA) of the 33K data, five of the most prominent axes of variation simply separated out the five breeds which were resequenced in the SNP discovery process (Gibbs et al. 2009). The discovery of the majority of SNPs within the 50K panel involved a next generation sequencing screen of only Holstein and Angus cattle, giving a heavy bias towards high minor allele frequencies within Northern European cattle (Van Tassell et al. 2008). These biases prevent reliable usage of a range of statistics which are based on the site frequency spectrum.

In this chapter, we investigated for the presence of positive selection using two approaches - one by identifying genetic divergence between populations, utilising the F_{ST} and LSBL tests, and another by determining if there is an excess of derived alleles in the population – both of which are more robust to the challenges above. We took an empirical approach to significance thresholds as ascertainment bias will likely be a strong confounder in using more complex models.

In both datasets F_{ST} was applied as a measure of divergence between continental populations, averaging the pairwise values where multiple European and Indian breeds were available. This was then examined in sliding windows of 5 SNPs and locus specific branch length (LSBL) was used to decompose values to the geographical groups. Designation of derived/ancestral status to alternate alleles was possible with sufficient reliability only with the 50K genotype data. Consequently, a genome-wide assessment of Fay and Wu's *H*, again averaged across 5 SNPs windows, was only performed on the Holstein, Somba and Indian zebu genotyped using that array.

The weaker of the two data sets, the 33K hapmap SNPs, offered the more limited analysis. Here, a lower density of SNPs coupled with an uneven distribution and the insecure calling of ancestral allele status rendered inference about adaptive signatures insecure. If one considers the top 5 ranked windows within each continental group, 7/15 of these are found on the three most densely sampled chromosomes (6, 14, 25) – reflecting, one suspects, a degree of signal noise in these data. Nevertheless, some genes may deserve some further consideration. The second ranked gene in African cattle, GPR149 has recently been shown to influence fertility levels in mice (Edson, Lin, and Matzuk 2010) domestic animal fertility is the subject of selection, both modern and presumably ancient. Also, within this continent, the 6th ranked window includes a gene coding for one of the major protein constituents of milk (CASA2); an obvious candidate for adaptive history within cattle.

The genes identified to fall within the top 1% of the LSBL windows were analysed in terms of Gene Ontology, in particular, if certain terms were over-represented. A relatively high number of categories were identified to be over-represented and with significant *P*-values. However, following multiple correction, the significant terms which remained associated (1) to the African population were those of protein binding and structural constituent of eye lens, (2) to the European population were that of keratin filament, and (3) to the Indicine population were those of cell adhesion and multicellular organismal development.

Among the significant terms, without multiple correction, were 5 categories of biological processes that were significant aross two populations, namely response to other organism, integrin-mediated signalling pathway, salivary gland development, visual perception and muscle thick filament assembly. None were found to be significant across all three populations. Two molecular function terms, however, were enriched across all three populations - protein binding and SH3/SH2 adaptor activity. A further nine molecular function terms were shared by two cattle populations, including chromatin binding, DNA-directed DNA polymerase activity, epidermal growth factor receptor activity and structural molecule activity.

The analysis of the 50K data represents a more promising examination of genome-wide bovine diversity. Within the African cattle population, on chromosome 3, several genes encoding guanylate binding proteins (GBP) were found within the top 15 contiguous windows from the LSBL analysis. These were also found to be within the top 1% of the ranked distribution in all three analyses. Members of this cluster include GBP3, GBP4, GBP5, GBP6 and GBP7. These genes are typically expressed following induction of immune-related genes. In addition to an association with the inflammatory response, some have also been postulated to be involved in cell proliferation (Olszewski, Gray, and Vestal 2006).

Also on chromosome 3, another set of contiguous SNP windows incorporates a series of genes of interest including melanophilin (MLPH) and prolactin releasing hormone (PRLH). They were detected using the LSBL (top 15 contiguous windows) and empirical composite analyses (top 1%). Melanophilin is a gene involved in pigmentation, and its mutation has been shown to affect the coat colour phenotype in dogs (Drogemuller et al. 2007) and cats (Ishida et al. 2006). Prolactin releasing hormone, on the other hand, is primarily associated with lactation, as implicated by its name (Hinuma et al. 1998). A positive selection signature for PRLH could conceivably be associated with dairying-related artifical selection within African domestic history.

On chromosome 10, a number of immune-related genes including casein kinase gamma (CSNK1G1), thyroid hormone receptor interactor 4 (TRIP4) and barrier to autointegration factor 1 (BANF1) were also detected across both the LSBL and Fay & Wu's *H* analyses.

Several other immune-related genes were identified among the most significant windows in the Fay & Wu's *H* analysis of the African cattle population. One of the top ranked windows contained CD40, which is tumour necrosis factor (TNF) receptor superfamily member 5. It plays a significant role in cattle immunity and is involved in the induction of interleukin-12 subunit p40 (IL-12p40) and inducible nitric oxide synthase (iNOS) genes. This response is suppressed in bovine monocyte-derived macrophages by *Mycobacterium avium* subspecies paratuberculosis (Sommer et al. 2009). CD40 also plays a vital role in the growth and differentiation of B-lymphocytes.

In the European population, a cluster of signalling lymphocytic activation molecule (SLAM) genes were found on chromosome 3 to display high inter-population differentiation. The cluster includes SLAMF1, CD84 (SLAMF5), SLAMF6, SLAMF7 and SLAMF9. Additionally, an interacting SLAM immunomodulator, lymphocyte antigen 9 (LY9) was also found. This is an important cluster of immune-related genes. Particularly, SLAMF1 has been shown to be implicated in Rinderpest, which is an infectious disease that affects cattle and where an outbreak brings significant economical ramifications. Wild type Rinderpest virus, a morbillovirus and a close relative of the human measles pathogen, requires SLAM as a receptor (Baron 2005). Rinderpest (literally, cattle plague) has a death rate in European cattle that often exceeds 80% and 200 million animals are estimated to have died in the eighteenth century alone (Huygelen 1997). It is possibly the strongest candidate infectious disease within bovines for having induced a major selective sweep due to selected differential susceptibility genomic variants.

Additionally, three contiguous window clusters among the top 15 from LSBL analysis also showed concordance with the top of 15 windows from the Fay & Wu's *H* analysis. The cluster of caveolin 1 (CAV1) and caveolin 2 (CAV2), along with testis derived transcript (TES) were found on chromosome 4. Caveolin are scaffolding proteins that are tumour suppressor gene candidates, and both CAV1 and CAV2 are required to form a stable complex together (Williams and Lisanti 2004). Testis derived transcript, on the other hand, is a testosterone responsive gene which encodes Sertoli cell secretory protein, highlighting its importance in reproduction.

Astrotactin 2 (ASTN2) and tripartite motif-containing 32 (TRIM32) on chromosome 8 were also detected across all European focused analyses as possibly under positive selection. ASTN2 is a neuronal adhesion molecule involved in glial-guided migration in cortical regions of the developing brain and olfactory bulb (Edmondson et al. 1988). TRIM32 functions to facilitate cell growth and its mutation have been implicated in muscular dystrophy (Cossee et al. 2009).

The third cluster contained a host of interesting developmental related genes, including pleiomorphic adenoma gene 1 (PLAG1) which has a salivary gland related function (Zhang et al. 2009), epidermal retinal dehydrogenase 2 (RDHE2) which affects height and skeletal frame size (Soranzo et al. 2009), and Moloney sarcoma oncogene (MOS) which plays a role in meiotic division of spermatocytes and oocytes (Cao et al. 2008; Prasad et al. 2008).

In the Indian cattle populations, diacylglycerol O-acyltransferase homolog 2 (DGAT2) located on chromosome X was also significantly ranked by LSBL as a candidate of positive selection. DGAT has previously been shown to be implicated in severe immunodeficiency conditions including psoriasis and diabetes (Yen et al. 2008). Within the same cluster as DGAT2 were a number of other genes of interest, including neuroligin3 (NLGN3) mutations which in humans are associated with autism and Asperger syndrome (Jamain et al. 2003); testis expressed 11 (TEX11), which is implicated in X-linked male infertility (Wang et al. 2001); and interleukin 2 receptor gamma (IL2RG), an important immune signalling component (Clark et al. 1995). This chromosome region also emerged within the 33K analysis top ranks.

The Fay & Wu's *H* analysis of the Indian population, however, unlike the European and the African populations, did not reveal concordance with the LSBL analysis of the same population among the top 15 contiguous windows.

Given that any individual test undoubtedly includes a high rate of false signals – particularly here where our data are limited to medium density coverage and have the complication of strong ascertainment bias – there is merit in comparing results across tests. This has a particular attraction where the tests are derived from relatively independent aspects of the data, as is the case with the two approaches used here.

Combining tests in a single statistic has been used effectively with divergence and extended haplotype homozygosity tests (Sabeti et al. 2007) and very recently Grossman *et al.* (Grossman et al. 2010) have proposed using a combination of multiple signals and argue that this gives greater precision as well as security in identifying adaptive variants. Here the F_{ST} and *H*-based empirical rankings of genomic windows are combined into a composite rank which is simply the product of the two individual ranks. The loci within these highest ranking regions are the strongest candidates in our study.

These composite empirical results revealed a measure of concordance between the LSBL and Fay & Wu's *H* analyses. In the African population, the GBP cluster once again topped the ranking. Similarly the cluster of CSNK1G1/TRIP4/BANF1 remained within the top 15 among the best ranked contiguous windows. However, the MLPH/PRLH cluster fell a little short, remaining within the top 1% of the distribution despite not being ranked among the top 15. We postulate these genes to be the most promising candidates for having undergone recent selective pressure, as the signals identified remain relatively strong through different statistical tests.

For the European cattle population, the CAV1/CAV2/TES cluster consistently ranks highly, followed closely by the clusters of PLAG1/RDHE2/MOS and TRIM32/ASTN2. These genes were highlighted in all three analyses as being among the top 15 contiguous windows.

Despite the lack of concordance between LSBL and Fay & Wu's H results for the Indian populations, the composite empirical analysis had some success in identifying genes with the best combined rank. DGAT2/TEX11/NLGN3/IL2RG was ranked highly, along with engrailed homeobox 1 (EN1), a gene involved in controlling development and pattern formation of the central nervous system.

A series of Gene Ontology analyses was also carried out. Similar to the outcome of the analyses carried out on the 33K SNP dataset, a relatively high number of categories were identified to be over-represented by genes identified to be putatively under positive selection. However, upon correction for multiple testing, these over-represented categories were no longer as significant. In one of the GO analyses, we combined all the genes which fell within the top 1% of the sliding windows of LSBL, Fay & Wu's *H* and composite analyses for each of the population for testing.

In terms of molecular functions, for categories significant prior to multiple testing, four terms were identified across all three populations of interest - nucleotide binding, NAD+ nucleosidase activity, potassium channel activity and protein binding. A further

18 terms were found to be over-represented across two populations, which include GTPase activity, voltage-gated potassium channel activity, calcium ion binding, lipoxygenase activity and transferase activity.

In terms of biological processes, also for those significant prior to multiple testing, two terms were over-represented in all three populations tested - potassium ion transport and NAD metabolic process. There were also 11 GO categories which are significant in two out of three populations, including ion transport, male meiosis, lipid metabolic process, positive regulation of B-cell proliferation and positive regulation of gamma-aminobutyric acid secretion.

Overall, our effort in detecting bovine genes under selection in geographically distinct population have met a certain degree of success, taking on two different statistical approaches – LSBL and Fay & Wu's *H* – and later combining them using a single composite statistic we are still able to identify outlying genes with relative concordance among the datasets.

The lack of concordance between results obtained using different approaches is not unexpected. Afterall, each method measures different "properties" of the dataset. F_{ST}, and by extension, LSBL, are used to measure divergence between populations. Fay & Wu's *H*, on the other hand, tests for excess of derived allele as well as events of selection that are of older time period in comparison to those tested using F_{ST}. The composite analysis aims to enable comparison between the two different measures by application of a ranking system, and to confer greater confidence of identifying adaptive variants.

SNPs that are associated with important traits could be studied in greater details in order to better understand the mechanism through which the polymorphism acts, and the impact the polymorphisms have on fitness of the population, in particular its effect on breeding and economic value of different cattle populations/breeds.

There are issues remaining with the 50K SNP dataset, chiefly the lack of SNP density and the presence of ascertainment bias in the set of SNPs genotyped. Ideally, a combination of the methods applied in this project thus far with methods such as extended haplotype homozygosity and site frequency spectrum analysis could be implemented, in order to provide more solid evidence of positive selection in the bovine genome.

A new high density 500K SNP chip from both Affymetrix and Illumina (currently in development) will work towards addressing the issue of SNP density. Additionally, the 500K chip will also have a broader SNP discovery process, hopefully reducing the problem of ascertainment bias. Ultimately, as technological advances are made and costs of utilising the technology fall, resequencing of entire genomes using next generation sequencing technology will be common, thus generating and providing dense, unbiased genome-wide datasets for more precise analyses in the search for signals of evolution in the genome. However, these and the other scans for the imprint of natural selection within the bovine genome presented in this thesis will require the corroboration of functional or epidemiological analyses for proof – there is a limit to the strength of inference that may be gained from observation of genetic variation alone. Within such analyses in humans, which in comparison represents a much more widely and densely sampled species, many inconclusive results with regard to selective signatures remain and may in fact never be resolved (Nielsen et al. 2007).

6. Conclusion

The primary aim of this thesis was to detect evidence of positive selection in mammalian genomes, and in particular, the bovine genome and in bovine populations. Over the years, advancement in technology and development of bioinformatics tools have enabled high throughput data generation and analyses at reasonable cost. Public databases have also served as repositories of genomic data where large amounts of information can be shared freely across the globe.

I began with a project to identify signatures of positive selection in the extracellular domains of human, chimpanzee and mouse orthologous genes, in preparation for applying this method to the bovine genome when it became available. Human and mouse, as model organisms, have been studied extensively and there was a wealth of information available, including completely sequenced genomes and validated genes. Past studies focused on detecting positive selection, primarily in humans, had usually estimated rates of change between non-synonymous and synonymous substitutions across entire genes which we hypothesised may be conservative and limiting. We postulated that it is likely that the occurrence of positive selection would be more frequent on the extracellular domain of a protein, given that the extracellular domain interacts directly with the external environment, including pathogens, ligands and other molecules that could drive selection.

Using orthologous gene trios, 13 genes with robust evidence for positive selection were detected. A pairwise comparative study between human and chimpanzee was also conducted to identify a further 16 genes, and our investigation showed that over full gene lengths, many of these candidate genes indeed contain a majority of their non-synonymous changes in the extracellular domain. The genes detected in this project include genes involved in immunity (CD34, CD37, ICAM2, TMC6), olfaction (OR1G1, OR4N5, OR5H1, OR6Y1, OR7G3), development (NTN4, ST8SIA3, ICAM5) and reproduction (SLC9A10, PEPP-2, ABHD1).

When GO analyses were performed on genes found with $d_N/d_S > 1$, we found enrichment of GO molecular function terms including receptor activity, olfactory receptor activity, melanocortin receptor activity, immunoglobulin E binding, complement receptor activity, chemokine activity, interleukin-10 receptor activity and G-protein coupled receptor activity. In terms of GO biological processes, the categories over-represented include sensory perception of smell, response to stimulus, immune response, chemotaxis, inflammatory response, amino acid transport, natural killer cell activation and cellular defense response. The ontology analyses support our hypothesis that the extracellular domain of immunity-related genes may be subject to positive selection due to their interactions with pathogens and exposure to external environments.

The extracellular domain protocol was further extended to identify genes under positive selection in cattle, by performing extracellular domain analyses on human, mouse, rat, dog and cow orthologous quintets. Two tests were carried out, one to compare the free-ratios model to the null model, and one to compare the bovine lineage specific model to the null model. A significant concordance between the two sets of results indicated that the method was a useful approach in the identification of genes under positive selection. Among the genes identified in these analyses were SLC26A8 and SLC2A5, which both affect spermatocyte development, and XLKD1 and SLAMF7, which have been reported to be a virulence factor and an immunomodulator, respectively.

Involvement with the Bovine Genome Sequencing and Analysis Consortium enabled performance of similar analyses on a newly curated bovine gene dataset, together with orthologues from human, mouse, rat, dog, platypus and opossum. However, not all orthologues were identified for each gene, necessitating analysis of each gene using independently reconstructed phylogenetic trees. As a result, not all generated phylogenies conform to the accepted general phylogeny of mammals. Nonetheless, in this study, genes were detected with signatures of positive selection in categories that have previously been reported to be subject to selection, including immune-related genes (IL24, IL15, IL23R, LEAP2, TREM1), genes involved in transcription (CBX7, HIST1H1C, ZNF771), lipid metabolism genes (FABP6, LIPE, PNPLA4) and cell adhesion genes (CD34).

Cross-projects results from analyses involving the extracellular domain method and orthologous genes across mammalian species show the enrichment in a number of gene categories that are involved in immunity, cell-cell adhesion,

The study of positive selection in this thesis was not limited to interspecies comparative analyses. The availability of ESTs and SNPs, as well as different statistical methods developed to analyse them, meant it was possible to investigate cattle populations for functionally relevant polymorphism and to search for signatures of selection at the population level in geographically separated breeds.

Before a large-scale SNP genotyping option was available, ESTs were used to detect potentially functionally relevant polymorphisms in bovine coding sequences. I attempted to identify synonymous and non-synonymous SNPs and using an algorithm known as SIFT, I generated a database of predicted tolerant or intolerant mutations. I also investigated for the presence of stop codon polymorphisms, which would have a dramatic effect on phenotype. Previously reported cases where loss of function mutations were maintained by positive selection include the stop codon polymorphism in human caspase-12, which confers resistance to severe sepsis, and in bovine myostatin, which gives rise to a double-muscling phenotype as well as improving the quality of beef cattle breeds. SOCS1 was identified in this study as a promising candidate following *in silico* analyses that provided strong evidence of the presence of a stop codon polymorphism. However, a sequencing screen in a panel of cattle samples failed to validate the presence of this mutation. I concluded that likelihood of dramatic stop codon polymorphisms is rare and the detection of these in ESTs is non-trivial.

This thesis also reports a contribution to the Bovine HapMap Consortium and investigating genotype data provided by the consortium for evidence of selection signatures in bovine breeds. The Bovine HapMap Consortium genotyped 501 animals sampled from 19 cattle breeds and 2 outgroups. F_{ST} statistics and a decomposed form, known as locus specific branch length (LSBL), were investigated by us to identify genomic regions with elevated genetic divergence in major bovine population groups (European *Bos taurus*, African *Bos taurus* and Indian *Bos indicus*). Geographically distinct breeds were shown to be subject to distinct selective pressures and several candidate genes likely under positive selection in these populations were identified. The 33K SNP data was, however, not without its limitations including low density, uneven SNP distribution across, with chromosomes 6, 14 and 25 most densely sampled, limited representation of *Indicus* and African breeds and ascertainment bias.

The subsequent availability of a 50K SNP chip enabled a more thorough study of positive selection in cattle populations. While the chip density remained relatively low, it addressed the issue of uneven sampling. Using bovine samples that were already available in our lab, bovine SNP genotyping was carried out with much better call rates in more geographically diverse breeds. F_{ST} and LSBL were used to detect genomic regions with elevated diversity between populations. I also took advantage of the successful genotyping of outgroup species to infer derived and ancestral allele frequencies, and were thus able to utilise Fay & Wu's *H*, a test for an excess of derived alleles indicative of positive selection. I also further extended the method to compare the two test statistics using an empirical composite method. The loci within highest ranking regions are the strongest candidates of positive selection.

In a European (Holstein-Friesian) cattle population, we noted a cluster of SLAM genes identified by high F_{ST} which are important in immunity. In particular, SLAMF1 has been previously shown to function as a receptor for rinderpest, an infectious disease associated with a high mortality rate during outbreaks. Note that SLAMF7 was also identified by inter-specific comparison (above). Within the African population (Somba), a large cluster of GBPs, which are involved in the inflammatory response, were consistently found ranked among the highest scores, Interesting candidate genes involved in pigmentation (MLPH) and lactation (PRLH) genes were also identified. For

Indian zebu, there was a lack of concordance between the different statistics. However, one cluster of genes DGAT2/ TEX11/ NLGN/ IL2RG, which have been implicated in immunodeficiency conditions, mental development, male infertility and immune signalling respectively, were consistently ranked highly in this population.

Gene Ontology analyses of 50K SNP dataset reveal enrichment of various GO molecular function terms of nucleotide binding, NAD+ nucleosidase activity, potassium channel activity, protein binding, GTPase activity, voltage-gated potassium channel activity, calcium ion binding, lipoxygenase activity and transferase activity. In terms of biological processes, significant terms include potassium ion transport, NAD metabolic process, ion transport, male meiosis, lipid metabolic process, positive regulation of B-cell proliferation and positive regulation of gamma-aminobutyric acid secretion.

Several methods were employed in the studies of this thesis, in order to interrogate the datasets that are both inter-species and intra-species in nature. The test of d_N/d_S was used to compare orthologous genes across different species to study the proportion of functional changes over time across millions of year. Fay & Wu's *H* was employed as statistical measure to test for high frequency derived alleles through inference of derived and ancestral allele. F_{ST} statistics was used to determine newer mutation and the differences in variants between populations of interest. Overall, the various studies carried out show that, across the board, genes involved in immunity, sensory of smell, protein binding, fertility and transporter activities are likely candidates for positive selection. These are gene categories that have often been shown, in various published studies, to be genes subject to positive selection.

In conclusion, this thesis reports a comprehensive investigation of positive selection in the bovine genome and in geographically diverse bovine populations featuring complementary approaches and different data sets. It also reports on investigation of positive selection in primate species.

Bibliography

Aaronson, J. S., B. Eckman, R. A. Blevins, J. A. Borkowski, J. Myerson, S. Imran, and K. O. Elliston. 1996. Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. Genome Res **6**:829-845.

Akey, J. M. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res **19**:711-722.

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res **12**:1805-1814.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol **215**:403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**:3389-3402.

Angata, K., M. Suzuki, J. McAuliffe, Y. Ding, O. Hindsgaul, and M. Fukuda. 2000. Differential biosynthesis of polysialic acid on neural cell adhesion molecule (NCAM) and oligosaccharide acceptors by three distinct alpha 2,8-sialyltransferases, ST8Sia IV (PST), ST8Sia II (STX), and ST8Sia III. J Biol Chem **275**:18594-18601.

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

Arbiza, L., J. Dopazo, and H. Dopazo. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol **2**:e38.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25**:25-29.

Bakewell, M. A., P. Shi, and J. Zhang. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci U S A **104**:7489-7494.

Bamshad, M., and S. P. Wooding. 2003. Signatures of natural selection in the human genome. Nat Rev Genet **4**:99-111.

Baron, M. D. 2005. Wild-type Rinderpest virus uses SLAM (CD150) as its receptor. J Gen Virol **86**:1753-1757.

Barone, S., S. L. Fussell, A. K. Singh, F. Lucas, J. Xu, C. Kim, X. Wu, Y. Yu, H. Amlal, U. Seidler, J. Zuo, and M. Soleimani. 2009. Slc2a5 (Glut5) is essential for the absorption of fructose in the intestine and generation of fructose-induced hypertension. J Biol Chem **284**:5056-5066.

Beja-Pereira, A., D. Caramelli, C. Lalueza-Fox, C. Vernesi, N. Ferrand, A. Casoli, F. Goyache, L. J. Royo, S. Conti, M. Lari, A. Martini, L. Ouragh, A. Magid, A. Atash, A. Zsolnai, P. Boscato, C. Triantaphylidis, K. Ploumi, L. Sineo, F. Mallegni, P. Taberlet, G. Erhardt, L. Sampietro, J. Bertranpetit, G. Barbujani, G. Luikart, and G. Bertorelle. 2006. The origin of European cattle: evidence from modern and ancient DNA. Proc Natl Acad Sci U S A **103**:8113-8118.

Beja-Pereira, A., G. Luikart, P. R. England, D. G. Bradley, O. C. Jann, G. Bertorelle, A. T. Chamberlain, T. P. Nunes, S. Metodiev, N. Ferrand, and G. Erhardt. 2003. Gene-culture coevolution between cattle milk protein genes and human lactase genes. Nat Genet **35**:311-313.

Bellinge, R. H., D. A. Liberles, S. P. Iaschi, A. O'Brien P, and G. K. Tay. 2005. Myostatin and its implications on animal breeding: a review. Anim Genet **36**:1-6.

Benjamini, Y., and Y. Hochberg. 1995a. Controlling the false discovery rate - a practice and powerful approach to multiple testing. J R Stat Soc Ser B **57**:289-300.

Benjamini, Y., and Y. Hochberg. 1995b. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological **57**:289-300.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2006. GenBank. Nucleic Acids Res **34**:D16-20.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet **74**:1111-1120.

Bertaux, C., and T. Dragic. 2006. Different domains of CD81 mediate distinct stages of hepatitis C virus pseudoparticle entry. J Virol **80**:4940-4948.

Blake-Palmer, K. G., Y. Su, A. N. Smith, and F. E. Karet. 2007. Molecular cloning and characterization of a novel form of the human vacuolar H+-ATPase e-subunit: an essential proton pump component. Gene **393**:94-100.

Boguski, M. S., T. M. Lowe, and C. M. Tolstoshev. 1993. dbEST--database for "expressed sequence tags". Nat Genet **4**:332-333.

Boguski, M. S., C. M. Tolstoshev, and D. E. Bassett, Jr. 1994. Gene discovery in dbEST. Science **265**:1993-1994.

Bradley, D. G., D. E. MacHugh, P. Cunningham, and R. T. Loftus. 1996. Mitochondrial diversity and the origins of African and European cattle. Proc Natl Acad Sci U S A **93**:5131-5135.

Burant, C. F., J. Takeda, E. Brot-Laroche, G. I. Bell, and N. O. Davidson. 1992. Fructose transporter in human spermatozoa and small intestine is GLUT5. J Biol Chem **267**:14523-14526.

Calonge, M. J., V. Volpini, L. Bisceglia, F. Rousaud, L. de Sanctis, E. Beccia, L. Zelante, X. Testar, A. Zorzano, X. Estivill, and et al. 1995. Genetic heterogeneity in cystinuria: the SLC3A1 gene is linked to type I but not to type III cystinuria. Proc Natl Acad Sci U S A **92**:9667-9671.

Cao, S. F., D. Li, Q. Yuan, X. Guan, and C. Xu. 2008. Spatial and temporal expression of c-mos in mouse testis during postnatal development. Asian J Androl **10**:277-285.

Chamary, J. V., J. L. Parmley, and L. D. Hurst. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet **7**:98-108.

Clark, A. G., S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302**:1960-1963.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen. 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res **15**:1496-1502.

Clark, P. A., T. Lester, S. Genet, A. M. Jones, R. Hendriks, R. J. Levinsky, and C. Kinnon. 1995. Screening for mutations causing X-linked severe combined immunodeficiency in the IL-2R gamma chain gene by single-strand conformation polymorphism analysis. Hum Genet **96**:427-432.

Cossee, M., C. Lagier-Tourenne, C. Seguela, M. Mohr, F. Leturcq, H. Gundesli, J. Chelly, C. Tranchant, M. Koenig, and J. L. Mandel. 2009. Use of SNP array analysis to identify a novel TRIM32 mutation in limb-girdle muscular dystrophy type 2H. Neuromuscul Disord **19**:255-260.

CSAC, C. S. a. A. C. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69-87.

Davis, S. J., S. Ikemizu, E. J. Evans, L. Fugger, T. R. Bakker, and P. A. van der Merwe. 2003. The nature of molecular recognition by T cells. Nat Immunol **4**:217-224.

Dean, M., M. Carrington, C. Winkler, G. A. Huttley, M. W. Smith, R. Allikmets, J. J. Goedert, S. P. Buchbinder, E. Vittinghoff, E. Gomperts, S. Donfield, D. Vlahov, R. Kaslow, A. Saah, C. Rinaldo, R. Detels, and S. J. O'Brien. 1996. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. Science **273**:1856-1862.

Drogemuller, C., U. Philipp, B. Haase, A. R. Gunzel-Apel, and T. Leeb. 2007. A noncoding melanophilin gene (MLPH) SNP at the splice donor of exon 1 represents a candidate causal mutation for coat color dilution in dogs. J Hered **98**:468-473.

Edgar, A. J. 2003. The gene structure and expression of human ABHD1: overlapping polyadenylation signal sequence with Sec12. BMC Genomics **4**:18.

Edmondson, J. C., R. K. Liem, J. E. Kuster, and M. E. Hatten. 1988. Astrotactin: a novel neuronal cell surface antigen that mediates neuron-astroglial interactions in cerebellar microcultures. J Cell Biol **106**:505-517.

Edson, M. A., Y. N. Lin, and M. M. Matzuk. 2010. Deletion of the novel oocyte-enriched gene, Gpr149, leads to increased fertility in mice. Endocrinology **151**:358-368.

Elsik, C. G., A. J. Mackey, J. T. Reese, N. V. Milshina, D. S. Roos, and G. M. Weinstock. 2007. Creating a honey bee consensus gene set. Genome Biol **8**:R13.

Elsik, C. G.R. L. TellamK. C. WorleyR. A. GibbsD. M. MuznyG. M. WeinstockD. L. AdelsonE. E. EichlerL. ElnitskiR. GuigoD. L. HamernikS. M. KappesH. A. LewinD. J. LynnF. W. NicholasA. ReymondM. RijnkelsL. C. SkowE. M. ZdobnovL. SchookJ. WomackT. AliotoS. E. AntonarakisA. AstashynC. E. ChappleH. C. ChenJ. ChrastF. CamaraO. ErmolaevaC. N. HenrichsenW. HlavinaY. KapustinB. KiryutinP. KittsF. KokocinskiM. LandrumD. MaglottK. PruittV. SapojnikovS. M. SearleV. SolovyevA. SouvorovC. UclaC. WyssJ. M. AnzolaD. GerlachE. ElhaikD. GraurJ. T. ReeseR. C. EdgarJ. C. McEwanG. M. PayneJ. M. RaisonT. JunierE. V. KriventsevaE. EyrasM. PlassR. DonthuD. M. LarkinJ. ReecyM. Q. YangL. ChenZ. ChengC. G. Chitko-McKownG. E. LiuL. K. MatukumalliJ. SongB. ZhuD. G. BradleyF. S. BrinkmanL. P. LauM. D. WhitesideA. WalkerT. T. WheelerT. CaseyJ. B. GermanD. G. LemayN. J. MaqboolA. J. MolenaarS. SeoP. StothardC. L. BaldwinR. BaxterC. L. Brinkmeyer-LangfordW. C. BrownC. P. ChildersT. ConnelleyS. A. EllisK. FritzE. J. GlassC. T. HerzigA. IivanainenK. K. LahmersA. K. BennettC. M. DickensJ. G. GilbertD. E. HagenH. SalihJ. AertsA. R. CaetanoB. DalrympleJ. F. GarciaC. A. GillS. G. HiendlederE. MemiliD. SpurlockJ. L. WilliamsL. AlexanderM. J. BrownsteinL. GuanR. A. HoltS. J. JonesM. A. MarraR. MooreS. S. MooreA. RobertsM. TaniguchiR. C. WatermanJ. ChackoM. M. ChandraboseA. CreeM. D. DaoH. H. DinhR. A. GabisiS. HinesJ. HumeS. N. JhangianiV. JoshiC. L. KovarL. R. LewisY. S. LiuJ. LopezM. B. MorganN. B. NguyenG. O. OkwuonuS. J. RuizJ. SantibanezR. A. WrightC. BuhayY. DingS. Dugan-RochaJ. HerdandezM. HolderA. SaboA. EganJ. GoodellK. Wilczek-BoneyG. R. FowlerM. E. HitchensR. J. LozadoC. MoenD. SteffenJ. T. WarrenJ.

ZhangR. ChiuJ. E. ScheinK. J. DurbinP. HavlakH. JiangY. LiuX. QinY. RenY. ShenH. SongS. N. BellC. DavisA. J. JohnsonS. LeeL. V. NazarethB. M. PatelL. L. PuS. VattathilR. L. Williams, Jr.S. CurryC. HamiltonE. SodergrenD. A. WheelerW. BarrisG. L. BennettA. EggenR. D. GreenG. P. HarhayM. HobbsO. JannJ. W. KeeleM. P. KentS. LienS. D. McKayS. McWilliamA. RatnakumarR. D. SchnabelT. SmithW. M. SnellingT. S. SonstegardR. T. StoneY. SugimotoA. TakasugaJ. F. TaylorC. P. Van TassellM. D. MacneilA. R. AbatepauloC. A. AbbeyV. AholaI. G. AlmeidaA. F. AmadioE. AnatrielloS. M. BahadueF. H. BiaseC. R. BoldtJ. A. CarrollW. A. CarvalhoE. P. CervelattiE. ChackoJ. E. ChapinY. ChengJ. ChoiA. J. ColleyT. A. de CamposM. De DonatoI. K. SantosC. J. de OliveiraH. DeobaldE. DevinoyK. E. DonohueP. DovcA. EberleinC. J. FitzsimmonsA. M. FranzinG. R. GarciaS. GeniniC. J. GladneyJ. R. GrantM. L. GreaserJ. A. GreenD. L. HadsellH. A. HakimovR. HalgrenJ. L. HarrowE. A. HartN. HastingsM. HernandezZ. L. HuA. InghamT. Iso-TouruC. JamisK. JensenD. KapetisT. KerrS. S. KhalilH. KhatibD. KolbehdariC. G. KumarD. KumarR. LeachJ. C. LeeC. LiK. M. LoganR. MalinverniE. MarquesW. F. MartinN. F. MartinsS. R. MaruyamaR. MazzaK. L. McLeanJ. F. MedranoB. T. MorenoD. D. MoreC. T. MunteanH. P. NandakumarM. F. NogueiraI. OlsakerS. D. PantF. PanzittaR. C. PastorM. A. PoliN. PoslusnyS. RachaganiS. RanganathanA. RazpetP. K. RiggsG. RinconN. Rodriguez-OsorioS. L. Rodriguez-ZasN. E. RomeroA. RosenwaldL. SandoS. M. SchmutzL. ShenL. ShermanB. R. SoutheyY. S. LutzowJ. V. SweedlerI. TammenB. P. TeluguJ. M. UrbanskiY. T. UtsunomiyaC. P. VerschoorA. J. WaardenbergZ. WangR. WardR. WeikardT. H. Welsh, Jr.S. N. WhiteL. G. WilmingK. R. WunderlichJ. Yang, and F. Q. Zhao. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science **324**:522-528.

Endo, T., K. Ikeo, and T. Gojobori. 1996. Large-scale search for genes on which positive selection may operate. Mol Biol Evol **13**:685-690.

Endo, T. A., M. Masuhara, M. Yokouchi, R. Suzuki, H. Sakamoto, K. Mitsui, A. Matsumoto, S. Tanimura, M. Ohtsubo, H. Misawa, T. Miyazaki, N. Leonor, T. Taniguchi, T. Fujita, Y. Kanakura, S. Komiya, and A. Yoshimura. 1997. A new protein containing an SH2 domain that inhibits JAK kinases. Nature **387**:921-924.

Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. Trends Ecol Evol **21**:569-575.

Fay, J. C., and C. I. Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics **155**:1405-1413.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ferrer-Costa, C., J. L. Gelpi, L. Zamakola, I. Parraga, X. de la Cruz, and M. Orozco. 2005. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics **21**:3176-3178.

Fitch, W. M., R. M. Bush, C. A. Bender, and N. J. Cox. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci U S A **94**:7712-7718.

Flori, L., S. Fritz, F. Jaffrezic, M. Boussaha, I. Gut, S. Heath, J. L. Foulley, and M. Gautier. 2009. The genome response to artificial selection: a case study in dairy cattle. PLoS One **4**:e6595.

Frazer, K. A.D. G. BallingerD. R. CoxD. A. HindsL. L. StuveR. A. GibbsJ. W. BelmontA. BoudreauP. HardenbolS. M. LealS. PasternakD. A. WheelerT. D. WillisF. YuH. YangC. ZengY. GaoH. HuW. HuC. LiW. LinS. LiuH. PanX. TangJ. WangW. WangJ. YuB. ZhangQ. ZhangH. ZhaoJ. ZhouS. B. GabrielR. BarryB. BlumenstielA. CamargoM. DefeliceM. FaggartM. GoyetteS. GuptaJ. MooreH. NguyenR. C. OnofrioM. ParkinJ. RoyE. StahlE. WinchesterL. ZiaugraD. AltshulerY. ShenZ. YaoW. HuangX. ChuY. HeL. JinY. LiuW. SunH. WangY. WangX. XiongL. XuM. M. WayeS. K. TsuiH. XueJ. T. WongL. M. GalverJ. B. FanK. GundersonS. S. MurrayA. R. OliphantM. S. CheeA. MontpetitF. ChagnonV. FerrettiM. LeboeufJ. F. OlivierM. S. PhillipsS. RoumyC. SalleeA. VernerT. J. HudsonP. Y. KwokD. CaiD. C. KoboldtR. D. MillerL. PawlikowskaP. Taillon-MillerM. XiaoL. C. TsuiW. MakY. Q. SongP. K. TamY. NakamuraT. KawaguchiT. KitamotoT. MorizonoA. NagashimaY. OhnishiA. SekineT. TanakaT. TsunodaP. DeloukasC. P. BirdM. DelgadoE. T. DermitzakisR. GwilliamS. HuntJ. MorrisonD. PowellB. E. StrangerP. WhittakerD. R. BentleyM. J. DalyP. I. de BakkerJ. BarrettY. R. ChretienJ. MallerS. McCarrollN. PattersonI. Pe'erA. PriceS. PurcellD. J. RichterP. SabetiR. SaxenaS. F. SchaffnerP. C. ShamP. VarillyL. D. SteinL. KrishnanA. V. SmithM. K. Tello-RuizG. A. ThorissonA. ChakravartiP. E. ChenD. J. CutlerC. S. KashukS. LinG. R. AbecasisW. GuanY. LiH. M. MunroZ. S. QinD. J. ThomasG. McVeanA. AutonL. BottoloN. CardinS. EyheramendyC. FreemanJ. MarchiniS. MyersC. SpencerM. StephensP. DonnellyL. R. CardonG. ClarkeD. M. EvansA. P. MorrisB. S. WeirJ. C. MullikinS. T. SherryM. FeoloA. SkolH. ZhangI. MatsudaY. FukushimaD. R. MacerE. SudaC. N. RotimiC. A. AdebamowoI. AjayiT. AniagwuP. A. MarshallC. NkwodimmahC. D. RoyalM. F. LeppertM. DixonA. PeifferR. QiuA. KentK. KatoN. NiikawaI. F. AdewoleB. M. KnoppersM. W. FosterE. W. ClaytonJ. WatkinD. MuznyL. NazarethE. SodergrenG. M. WeinstockI. YakubB. W. BirrenR. K. WilsonL. L. FultonJ. RogersJ. BurtonN. P. CarterC. M. CleeM. GriffithsM. C. JonesK. McLayR. W. PlumbM. T. RossS. K. SimsD. L. WilleyZ. ChenH. HanL. KangM. GodboutJ. C. WallenburgP. L'ArchevequeG. BellemareK. SaekiD. AnH. FuQ. LiZ. WangR. WangA. L. HoldenL. D. BrooksJ. E. McEwenM. S. GuyerV. O. WangJ. L. PetersonM. ShiJ. SpiegelL. M. SungL. F. ZachariaF. S. CollinsK. KennedyR. Jamieson, and J. Stewart. 2007a. A second generation human haplotype map of over 3.1 million SNPs. Nature **449**:851-861.

Frazer, K. A., E. Eskin, H. M. Kang, M. A. Bogue, D. A. Hinds, E. J. Beilharz, R. V. Gupta, J. Montgomery, M. M. Morenzoni, G. B. Nilsen, C. L. Pethiyagoda, L. L. Stuve, F. M. Johnson, M. J. Daly, C. M. Wade, and D. R. Cox. 2007b. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature **448**:1050-1053.

Freeman, A. R., D. J. Lynn, C. Murray, and D. G. Bradley. 2008. Detecting the effects of selection at the population level in six bovine immune genes. BMC Genet **9**:62.

Fu, Y. X., and W. H. Li. 1993. Statistical tests of neutrality of mutations. Genetics **133**:693-709.

Gagneux, P., and A. Varki. 2001. Genetic differences between humans and great apes. Mol Phylogenet Evol **18**:2-13.

Gibbs, R. A.J. RogersM. G. KatzeR. BumgarnerG. M. WeinstockE. R. MardisK. A. RemingtonR. L. StrausbergJ. C. VenterR. K. WilsonM. A. BatzerC. D. BustamanteE. E. EichlerM. W. HahnR. C. HardisonK. D. MakovaW. MillerA. MilosavljevicR. E. PalermoA. SiepelJ. M. SikelaT. AttawayS. BellK. E. BernardC. J. BuhayM. N. ChandraboseM. DaoC. DavisK. D. DelehauntyY. DingH. H. DinhS. Dugan-RochaL. A. FultonR. A. GabisiT. T. GarnerJ. GodfreyA. C. HawesJ. HernandezS. HinesM. HolderJ. HumeS. N. JhangianiV. JoshiZ. M. KhanE. F. KirknessA. CreeR. G. FowlerS. LeeL. R. LewisZ. LiY. S. LiuS. M. MooreD. MuznyL. V. NazarethD. N. NgoG. O. OkwuonuG. PaiD. ParkerH. A. PaulC. PfannkochC. S. PohlY. H. RogersS. J. RuizA. SaboJ. SantibanezB. W. SchneiderS. M. SmithE. SodergrenA. F. SvatekT. R. UtterbackS. VattathilW. WarrenC. S. WhiteA. T. ChinwallaY. FengA. L. HalpernL. W. HillierX. HuangP. MinxJ. O. NelsonK. H. PepinX. QinG. G. SuttonE. VenterB. P. WalenzJ. W. WallisK. C. WorleyS. P. YangS. M. JonesM. A. MarraM. RocchiJ. E. ScheinR. BaertschL. ClarkeM. CsurosJ. GlasscockR. A. HarrisP. HavlakA. R. JacksonH. JiangY. LiuD. N. MessinaY. ShenH. X. SongT. WylieL. ZhangE. BirneyK. HanM. K. KonkelJ. LeeA. F. SmitB. UllmerH. WangJ. XingR. BurhansZ. ChengJ. E. KarroJ. MaB. RaneyX. SheM. J. CoxJ. P. DemuthL. J. DumasS. G. HanJ. HopkinsA. Karimpour-FardY. H. KimJ. R. PollackT. VinarC. Addo-QuayeJ. DegenhardtA. DenbyM. J. HubiszA. IndapC. KosiolB. T. LahnH. A. LawsonA. MarkleinR. NielsenE. J. VallenderA. G. ClarkB. FergusonR. D. HernandezK. HiraniH. Kehrer-SawatzkiJ. KolbS. PatilL. L. PuY. RenD. G. SmithD. A. WheelerI. SchenckE. V. BallR. ChenD. N. CooperB. GiardineF. HsuW. J. KentA. LeskD. L. NelsonE. O'Brien WK. PruferP. D. StensonJ. C. WallaceH. KeX. M. LiuP. WangA. P. XiangF. YangG. P. BarberD. HausslerD. KarolchikA. D. KernR. M. KuhnK. E. Smith, and A. S. Zwieg. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science **316**:222-234.

Gibbs, R. A., J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole, C. A. Gill, R. D. Green, D. L. Hamernik, S. M. Kappes, S. Lien, L. K. Matukumalli, J. C. McEwan, L. V. Nazareth, R. D. Schnabel, G. M. Weinstock, D. A. Wheeler, P. Ajmone-Marsan, P. J. Boettcher, A. R. Caetano, J. F. Garcia, O. Hanotte, P. Mariani, L. C. Skow, T. S. Sonstegard, J. L. Williams, B. Diallo, L. Hailemariam, M. L. Martinez, C. A. Morris, L. O. Silva, R. J. Spelman, W. Mulatu, K. Zhao, C. A. Abbey, M. Agaba, F. R. Araujo, R. J. Bunch, J. Burton, C. Gorni, H. Olivier, B. E. Harrison, B. Luff, M. A. Machado, J. Mwakaya, G. Plastow, W. Sim, T. Smith, M. B. Thomas, A. Valentini, P. Williams, J. Womack, J. A. Woolliams, Y. Liu, X. Qin, K. C. Worley, C. Gao, H. Jiang, S. S. Moore, Y. Ren, X. Z. Song, C. D. Bustamante, R. D. Hernandez, D. M. Muzny, S. Patil, A. San Lucas, Q. Fu, M. P. Kent, R. Vega, A.

Matukumalli, S. McWilliam, G. Sclep, K. Bryc, J. Choi, H. Gao, J. J. Grefenstette, B. Murdoch, A. Stella, R. Villa-Angulo, M. Wright, J. Aerts, O. Jann, R. Negrini, M. E. Goddard, B. J. Hayes, D. G. Bradley, M. Barbosa da Silva, L. P. Lau, G. E. Liu, D. J. Lynn, F. Panzitta, and K. G. Dodds. 2005. A haplotype map of the human genome. Nature **437**:1299-1320.

Gibbs, R. A., J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole, C. A. Gill, R. D. Green, D. L. Hamernik, S. M. Kappes, S. Lien, L. K. Matukumalli, J. C. McEwan, L. V. Nazareth, R. D. Schnabel, G. M. Weinstock, D. A. Wheeler, P. Ajmone-Marsan, P. J. Boettcher, A. R. Caetano, J. F. Garcia, O. Hanotte, P. Mariani, L. C. Skow, T. S. Sonstegard, J. L. Williams, B. Diallo, L. Hailemariam, M. L. Martinez, C. A. Morris, L. O. Silva, R. J. Spelman, W. Mulatu, K. Zhao, C. A. Abbey, M. Agaba, F. R. Araujo, R. J. Bunch, J. Burton, C. Gorni, H. Olivier, B. E. Harrison, B. Luff, M. A. Machado, J. Mwakaya, G. Plastow, W. Sim, T. Smith, M. B. Thomas, A. Valentini, P. Williams, J. Womack, J. A. Woolliams, Y. Liu, X. Qin, K. C. Worley, C. Gao, H. Jiang, S. S. Moore, Y. Ren, X. Z. Song, C. D. Bustamante, R. D. Hernandez, D. M. Muzny, S. Patil, A. San Lucas, Q. Fu, M. P. Kent, R. Vega, A. Matukumalli, S. McWilliam, G. Sclep, K. Bryc, J. Choi, H. Gao, J. J. Grefenstette, B. Murdoch, A. Stella, R. Villa-Angulo, M. Wright, J. Aerts, O. Jann, R. Negrini, M. E. Goddard, B. J. Hayes, D. G. Bradley, M. Barbosa da Silva, L. P. Lau, G. E. Liu, D. J. Lynn, F. Panzitta, and K. G. Dodds. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science **324**:528-532.

Gibbs, R. A.G. M. WeinstockM. L. MetzkerD. M. MuznyE. J. SodergrenS. SchererG. ScottD. SteffenK. C. WorleyP. E. BurchG. OkwuonuS. HinesL. LewisC. DeRamoO. DelgadoS. Dugan-RochaG. MinerM. MorganA. HawesR. GillCeleraR. A. HoltM. D. AdamsP. G. AmanatidesH. Baden-TillsonM. BarnsteadS. ChinC. A. EvansS. FerrieraC. FoslerA. GlodekZ. GuD. JenningsC. L. KraftT. NguyenC. M. PfannkochC. SitterG. G. SuttonJ. C. VenterT. WoodageD. SmithH. M. LeeE. GustafsonP. CahillA. KanaL. Doucette-StammK. WeinstockK. FechtelR. B. WeissD. M. DunnE. D. GreenR. W. BlakesleyG. G. BouffardP. J. De JongK. OsoegawaB. ZhuM. MarraJ. ScheinI. BosdetC. FjellS. JonesM. KrzywinskiC. MathewsonA. SiddiquiN. WyeJ. McPhersonS. ZhaoC. M. FraserJ. ShettyS. ShatsmanK. GeerY. ChenS. AbramzonW. C. NiermanP. H. HavlakR. ChenK. J. DurbinA. EganY. RenX. Z. SongB. LiY. LiuX. QinS. CawleyA. J. CooneyL. M. D'SouzaK. MartinJ. Q. WuM. L. Gonzalez-GarayA. R. JacksonK. J. KalafusM. P. McLeodA. MilosavljevicD. VirkA. VolkovD. A. WheelerZ. ZhangJ. A. BaileyE. E. EichlerE. TuzunE. BirneyE. MonginA. Ureta-VidalC. WoodwarkE. ZdobnovP. BorkM. SuyamaD. TorrentsM. AlexanderssonB. J. TraskJ. M. YoungH. HuangH. WangH. XingS. DanielsD. GietzenJ. SchmidtK. StevensU. VittJ. WingroveF. CamaraM. Mar AlbaJ. F. AbrilR. GuigoA. SmitI. DubchakE. M. RubinO. CouronneA. PoliakovN. HubnerD. GantenC. GoeseleO. HummelT. KreitlerY. A. LeeJ. MontiH. SchulzH. ZimdahlH. HimmelbauerH. LehrachH. J. JacobS. BrombergJ. Gullings-HandleyM. I. Jensen-SeamanA. E. KwitekJ. LazarD. PaskoP. J. TonellatoS. TwiggerC. P. PontingJ. M. DuarteS. RiceL. GoodstadtS. A. BeatsonR. D. EmesE. E. WinterC. WebberP. BrandtG. NyakaturaM. AdetobiF. ChiaromonteL. ElnitskiP. EswaraR. C. HardisonM. HouD. KolbeK. MakovaW.

MillerA. NekrutenkoC. RiemerS. SchwartzJ. TaylorS. YangY. ZhangK. LindpaintnerT. D. AndrewsM. CaccamoM. ClampL. ClarkeV. CurwenR. DurbinE. EyrasS. M. SearleG. M. CooperS. BatzoglouM. BrudnoA. SidowE. A. StoneB. A. PayseurG. BourqueC. Lopez-OtinX. S. PuenteK. ChakrabartiS. ChatterjiC. DeweyL. PachterN. BrayV. B. YapA. CaspiG. TeslerP. A. PevznerD. HausslerK. M. RoskinR. BaertschH. ClawsonT. S. FureyA. S. HinrichsD. KarolchikW. J. KentK. R. RosenbloomH. TrumbowerM. WeirauchD. N. CooperP. D. StensonB. MaM. BrentM. ArumugamD. ShteynbergR. R. CopleyM. S. TaylorH. RiethmanU. MudunuriJ. PetersonM. GuyerA. FelsenfeldS. OldS. Mockrin, and F. Collins. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**:493-521.

Gilad, Y., C. D. Bustamante, D. Lancet, and S. Paabo. 2003. Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet **73**:489-501.

Gilad, Y., O. Man, and G. Glusman. 2005. A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Res **15**:224-230.

Gilad, Y., A. Oshlack, G. K. Smyth, T. P. Speed, and K. P. White. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature **440**:242-245.

Grossman, S. R., I. Shylakhter, E. K. Karlsson, E. H. Byrne, S. Morales, G. Frieden, E. Hostetter, E. Angelino, M. Garber, O. Zuk, E. S. Lander, S. F. Schaffner, and P. C. Sabeti. 2010. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science **epub ahead of print citation**.

Hanel, K., T. Stangler, M. Stoldt, and D. Willbold. 2006. Solution structure of the X4 protein coded by the SARS related coronavirus reveals an immunoglobulin like fold and suggests a binding activity to integrin I domains. J Biomed Sci **13**:281-293.

Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res **32**:D258-261.

Hawken, R. J., W. C. Barris, S. M. McWilliam, and B. P. Dalrymple. 2004. An interactive bovine in silico SNP database (IBISS). Mamm Genome **15**:819-827.

Heaton, M. P., G. P. Harhay, G. L. Bennett, R. T. Stone, W. M. Grosse, E. Casas, J. W. Keele, T. P. Smith, C. G. Chitko-McKown, and W. W. Laegreid. 2002. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. Mamm Genome **13**:272-281.

Heaton, M. P., J. E. Keen, M. L. Clawson, G. P. Harhay, N. Bauer, C. Shultz, B. T. Green, L. Durso, C. G. Chitko-McKown, and W. W. Laegreid. 2005. Use of bovine single nucleotide polymorphism markers to verify sample tracking in beef processing. J Am Vet Med Assoc **226**:1311-1314.

Hedrick, P. W., and B. C. Verrelli. 2006. "Ground truth" for selection on CCR5-Delta32. Trends Genet **22**:293-296.

Helmer, D., L. Gourichon, H. Monchot, J. Peters, and M. Sana Segui. 2005. Identifying early domestica cattle from Pre-pottery Neolithic sites on the Middle Euphrates using sexual dimorphism. . Pp. 86-95 in J. P. J-D Vigne, D Helmer., ed. The first steps of animal domestication: New arcaheological approaches. Oxbow Books, Oxford.

Hinuma, S., Y. Habata, R. Fujii, Y. Kawamata, M. Hosoya, S. Fukusumi, C. Kitada, Y. Masuo, T. Asano, H. Matsumoto, M. Sekiguchi, T. Kurokawa, O. Nishimura, H. Onda, and M. Fujino. 1998. A prolactin-releasing peptide in the brain. Nature **393**:272-276.

Hobolth, A., O. F. Christensen, T. Mailund, and M. H. Schierup. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet **3**:e7.

Horejsi, V., and C. Vlcek. 1991. Novel structurally distinct family of leucocyte surface glycoproteins including CD9, CD37, CD53 and CD63. FEBS Lett **288**:1-4.

Hulo, N., A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. Sigrist. 2006. The PROSITE database. Nucleic Acids Res **34**:D227-230.

Huygelen, C. 1997. The immunization of cattle against rinderpest in eighteenth-century Europe. Med Hist **41**:182-196.

Ishida, Y., V. A. David, E. Eizirik, A. A. Schaffer, B. A. Neelam, M. E. Roelke, S. S. Hannah, J. O'Brien S, and M. Menotti-Raymond. 2006. A homozygous single-base deletion in MLPH causes the dilute coat color phenotype in the domestic cat. Genomics **88**:698-705.

Jamain, S., H. Quach, C. Betancur, M. Rastam, C. Colineaux, I. C. Gillberg, H. Soderstrom, B. Giros, M. Leboyer, C. Gillberg, and T. Bourgeron. 2003. Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. Nat Genet **34**:27-29.

Karlsson, E. K., I. Baranowska, C. M. Wade, N. H. Salmon Hillbertz, M. C. Zody, N. Anderson, T. M. Biagi, N. Patterson, G. R. Pielberg, E. J. Kulbokas, 3rd, K. E. Comstock, E. T. Keller, J. P. Mesirov, H. von Euler, O. Kampe, A. Hedhammar, E. S. Lander, G. Andersson, L. Andersson, and K. Lindblad-Toh. 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet **39**:1321-1328.

Kent, W. J. 2002. BLAT--the BLAST-like alignment tool. Genome Res **12**:656-664.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science **309**:1850-1854.

Kijimoto-Ochiai, S. 2002. CD23 (the low-affinity IgE receptor) as a C-type lectin: a multidomain and multifunctional molecule. Cell Mol Life Sci **59**:648-664.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature **217**:624-626.

Koch, M., J. R. Murrell, D. D. Hunter, P. F. Olson, W. Jin, D. R. Keene, W. J. Brunken, and R. E. Burgeson. 2000. A novel member of the netrin family, beta-netrin, shares homology with the beta chain of laminin: identification, expression, and functional characterization. J Cell Biol **151**:221-234.

Kosiol, C., T. Vinar, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen, and A. Siepel. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet **4**:e1000144.

Krawczak, M., E. V. Ball, I. Fenton, P. D. Stenson, S. Abeysinghe, N. Thomas, and D. N. Cooper. 2000. Human gene mutation database-a biomedical information and research resource. Hum Mutat **15**:45-51.

Kreitman, M. 1996. The neutral theory is dead. Long live the neutral theory. Bioessays **18**:678-683; discussion 683.

Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol **305**:567-580.

Kumar, P., S. Henikoff, and P. C. Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc **4**:1073-1081.

Lamason, R. L., M. A. Mohideen, J. R. Mest, A. C. Wong, H. L. Norton, M. C. Aros, M. J. Jurynec, X. Mao, V. R. Humphreville, J. E. Humbert, S. Sinha, J. L. Moore, P. Jagadeeswaran, W. Zhao, G. Ning, I. Makalowska, P. M. McKeigue, D. O'Donnell, R. Kittles, E. J. Parra, N. J. Mangini, D. J. Grunwald, M. D. Shriver, V. A. Canfield, and K. C. Cheng. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science **310**:1782-1786.

Lander, E. S.L. M. LintonB. BirrenC. NusbaumM. C. ZodyJ. BaldwinK. DevonK. DewarM. DoyleW. FitzHughR. FunkeD. GageK. HarrisA. HeafordJ. HowlandL. KannJ. LehoczkyR. LeVineP. McEwanK. McKernanJ. MeldrimJ. P. MesirovC. MirandaW. MorrisJ. NaylorC. RaymondM. RosettiR. SantosA. SheridanC. SougnezN. Stange-ThomannN. StojanovicA. SubramanianD. WymanJ. RogersJ. SulstonR. AinscoughS. BeckD. BentleyJ. BurtonC. CleeN. CarterA. CoulsonR. DeadmanP. DeloukasA. DunhamI. DunhamR. DurbinL. FrenchD. GrafhamS. GregoryT. HubbardS. HumphrayA. HuntM. JonesC. LloydA. McMurrayL. MatthewsS. MercerS. MilneJ. C. MullikinA. MungallR. PlumbM. RossR. ShownkeenS. SimsR. H. WaterstonR. K. WilsonL. W. HillierJ. D. McPhersonM. A. MarraE. R. MardisL. A. FultonA. T. ChinwallaK. H. PepinW. R. GishS. L. ChissoeM. C. WendlK. D. DelehauntyT. L. MinerA. DelehauntyJ. B. KramerL. L. CookR. S. FultonD. L. JohnsonP. J. MinxS. W. CliftonT. HawkinsE. BranscombP. PredkiP. RichardsonS. WenningT. SlezakN. DoggettJ. F. ChengA. OlsenS. LucasC. ElkinE. UberbacherM. FrazierR. A. GibbsD. M. MuznyS. E. SchererJ. B. BouckE. J. SodergrenK. C. WorleyC. M. RivesJ. H. GorrellM. L. MetzkerS. L. NaylorR. S. KucherlapatiD. L. NelsonG. M. WeinstockY. SakakiA. FujiyamaM. HattoriT. YadaA. ToyodaT. ItohC. KawagoeH. WatanabeY. TotokiT. TaylorJ. WeissenbachR. HeiligW. SaurinF. ArtiguenaveP. BrottierT. BrulsE. PelletierC. RobertP. WinckerD. R. SmithL. Doucette-StammM. RubenfieldK. WeinstockH. M. LeeJ. DuboisA. RosenthalM. PlatzerG. NyakaturaS. TaudienA. RumpH. YangJ. YuJ. WangG. HuangJ. GuL. HoodL. RowenA. MadanS. QinR. W. DavisN. A. FederspielA. P. AbolaM. J. ProctorR. M. MyersJ. SchmutzM. DicksonJ. GrimwoodD. R. CoxM. V. OlsonR. KaulN. ShimizuK. KawasakiS. MinoshimaG. A. EvansM. AthanasiouR. SchultzB. A. RoeF. ChenH. PanJ. RamserH. LehrachR. ReinhardtW. R. McCombieM. de la BastideN. DedhiaH. BlockerK. HornischerG. NordsiekR. AgarwalaL. AravindJ. A. BaileyA. BatemanS. BatzoglouE. BirneyP. BorkD. G. BrownC. B. BurgeL. CeruttiH. C. ChenD. ChurchM. ClampR. R. CopleyT. DoerksS. R. EddyE. E. EichlerT. S. FureyJ. GalaganJ. G. GilbertC. HarmonY. HayashizakiD. HausslerH. HermjakobK. HokampW. JangL. S. JohnsonT. A. JonesS. KasifA. KaspryzkS. KennedyW. J. KentP. KittsE. V. KooninI. KorfD. KulpD. LancetT. M. LoweA. McLysaghtT. MikkelsenJ. V. MoranN. MulderV. J. PollaraC. P. PontingG. SchulerJ. SchultzG. SlaterA. F. SmitE. StupkaJ. SzustakowskiD. Thierry-MiegJ. Thierry-MiegL. WagnerJ. WallisR. WheelerA. WilliamsY. I. WolfK. H. WolfeS. P. YangR. F. YehF. CollinsM. S. GuyerJ. PetersonA. FelsenfeldK. A. WetterstrandA. PatrinosM. J. MorganP. de JongJ. J. CataneseK. OsoegawaH. ShizuyaS. Choi, and Y. J. Chen. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Lao, D. M., T. Okuno, and T. Shimizu. 2002. Evaluating transmembrane topology prediction methods for the effect of signal peptide in topology prediction. In Silico Biol **2**:485-494.

Larsen, L., and C. Ropke. 2002. Suppressors of cytokine signalling: SOCS. APMIS **110**:833-844.

Lewontin, R. C., and J. L. Hubby. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics **54**:595-609.

Lhuillier, P., B. Rode, D. Escalier, P. Lores, T. Dirami, T. Bienvenu, G. Gacon, E. Dulioust, and A. Toure. 2009. Absence of annulus in human asthenozoospermia: case report. Hum Reprod **24**:1296-1303.

Li, W. H., and M. A. Saunders. 2005. News and views: the chimpanzee and us. Nature **437**:50-51.

Liang, Z. G., P. A. O'Hern, B. Yavetz, H. Yavetz, and E. Goldberg. 1994. Human testis cDNAs identified by sera from infertile patients: a molecular biological approach to immunocontraceptive development. Reprod Fertil Dev **6**:297-305.

Lindblad-Toh, K.C. M. WadeT. S. MikkelsenE. K. KarlssonD. B. JaffeM. KamalM. ClampJ. L. ChangE. J. Kulbokas, 3rdM. C. ZodyE. MauceliX. XieM. BreenR. K. WayneE. A. OstranderC. P. PontingF. GalibertD. R. SmithP. J. DeJongE. KirknessP. AlvarezT. BiagiW. BrockmanJ. ButlerC. W. ChinA. CookJ. CuffM. J. DalyD. DeCaprioS. GnerreM. GrabherrM. KellisM. KleberC. BardelebenL. GoodstadtA. HegerC. HitteL. KimK. P. KoepfliH. G. ParkerJ. P. PollingerS. M. SearleN. B. SutterR. ThomasC. WebberJ. BaldwinA. AbebeA. AbouelleilL. AftuckM. Ait-ZahraT. AldredgeN. AllenP. AnS. AndersonC. AntoineH. ArachchiA. AslamL. AyotteP. BachantsangA. BarryT. BayulM. BenamaraA. BerlinD. BessetteB. BlitshteynT. BloomJ. BlyeL. BoguslavskiyC. BonnetB. BoukhgalterA. BrownP. CahillN. CalixteJ. CamarataY. CheshatsangJ. ChuM. CitroenA. CollymoreP. CookeT. DawoeR. DazaK. DecktorS. DeGrayN. DhargayK. DooleyP. DorjeK. DorjeeL. DorrisN. DuffeyA. DupesO. EgbiremolenR. ElongJ. FalkA. FarinaS. FaroD. FergusonP. FerreiraS. FisherM. FitzGeraldK. FoleyC. FoleyA. FrankeD. FriedrichD. GageM. GarberG. GearinG. GiannoukosT. GoodeA. GoyetteJ. GrahamE. GrandboisK. GyaltsenN. HafezD. HagopianB. HagosJ. HallC. HealyR. HegartyT. HonanA. HornN. HoudeL. HughesL. HunnicuttM. HusbyB. JesterC. JonesA. KamatB. KangaC. KellsD. KhazanovichA. C. KieuP. KisnerM. KumarK. LanceT. LandersM. LaraW. LeeJ. P. LegerN. LennonL. LeuperS. LeVineJ. LiuX. LiuY. LokyitsangT. LokyitsangA. LuiJ. MacdonaldJ. MajorR. MarabellaK. MaruC. MatthewsS. McDonoughT. MehtaJ. MeldrimA. MelnikovL. MeneusA. MihalevT. MihovaK. MillerR. MittelmanV. MlengaL. MulrainG. MunsonA. NavidiJ. NaylorT. NguyenN. NguyenC. NguyenR. NicolN. NorbuC. NorbuN. NovodT. NyimaP. OlandtB. O'NeillK. O'NeillS. OsmanL. OyonoC. PattiD. PerrinP. PhunkhangF. PierreM. PriestA. RachupkaS. RaghuramanR. RameauV. RayC. RaymondF. RegeC. RiseJ. RogersP. RogovJ. SahalieS. SettipalliT. SharpeT. SheaM. SheehanN. SherpaJ. ShiD. ShihJ. SloanC. SmithT. SparrowJ. StalkerN. Stange-ThomannS. StavropoulosC. StoneS. StoneS. SykesP. TchuingaP. TenzingS. TesfayeD. ThoulutsangY. ThoulutsangK. TophamI. ToppingT. TsamlaH. VassilievV. VenkataramanA. VoT. WangchukT. WangdiM. WeiandJ. WilkinsonA. WilsonS. YadavS. YangX. YangG. YoungQ. YuJ. ZainounL. ZembekA. Zimmer, and E. S.

Lander. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature **438**:803-819.

Lindqvist, P. G., P. J. Svensson, B. Dahlback, and K. Marsal. 1998. Factor V Q506 mutation (activated protein C resistance) associated with reduced intrapartum blood loss--a possible evolutionary selection mechanism. Thromb Haemost **79**:69-73.

Loftus, R. T., D. E. MacHugh, D. G. Bradley, P. M. Sharp, and P. Cunningham. 1994. Evidence for two independent domestications of cattle. Proc Natl Acad Sci U S A **91**:2757-2761.

Losman, J. A., X. P. Chen, D. Hilton, and P. Rothman. 1999. Cutting edge: SOCS-1 is a potent inhibitor of IL-4 signal transduction. J Immunol **162**:3770-3774.

Lusso, P. 2006. HIV and the chemokine system: 10 years later. EMBO J **25**:447-456.

Lynn, D. J., A. R. Freeman, C. Murray, and D. G. Bradley. 2005. A genomics approach to the detection of positive selection in cattle: adaptive evolution of the T-cell and natural killer cell-surface protein CD2. Genetics **170**:1189-1196.

Lynn, D. J., G. L. Winsor, C. Chan, N. Richard, M. R. Laird, A. Barsky, J. L. Gardy, F. M. Roche, T. H. Chan, N. Shah, R. Lo, M. Naseer, J. Que, M. Yau, M. Acab, D. Tulpan, M. D. Whiteside, A. Chikatamarla, B. Mah, T. Munzner, K. Hokamp, R. E. Hancock, and F. S. Brinkman. 2008a. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Mol Syst Biol **4**:218.

Lynn, D. J., G. L. Winsor, C. Chan, N. Richard, M. R. Laird, A. Barsky, J. L. Gardy, F. M. Roche, T. H. W. Chan, N. Shah, R. Lo, M. Naseer, J. Que, M. Yau, M. Acab, D. Tulpan, M. Whiteside, A. Chikatamarla, B. Mah, T. M. Munzner, K. Hokamp, R. E. W. Hancock, and F. S. L. Brinkman. 2008b. Facilitating systems biology approaches to studying mammalian innate immunity.

MacEachern, S., B. Hayes, J. McEwan, and M. Goddard. 2009. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics **10**:181.

MacEachern, S., J. McEwan, and M. Goddard. 2009. Phylogenetic reconstruction and the identification of ancient polymorphism in the Bovini tribe (Bovidae, Bovinae). BMC Genomics **10**:177.

Makela, S., R. Eklund, J. Lahdetie, M. Mikkola, O. Hovatta, and J. Kere. 2005. Mutational analysis of the human SLC26A8 gene: exclusion as a candidate for male infertility due to primary spermatogenic failure. Mol Hum Reprod **11**:129-132.

Malhi, R. S., B. Sickler, D. Lin, J. Satkoski, R. Y. Tito, D. George, S. Kanthaswamy, and D. G. Smith. 2007. MamuSNP: a resource for Rhesus Macaque (Macaca mulatta) genomics. PLoS One **2**:e438.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One **4**:e5350.

McLatchie, L. M., N. J. Fraser, M. J. Main, A. Wise, J. Brown, N. Thompson, R. Solari, M. G. Lee, and S. M. Foord. 1998. RAMPs regulate the transport and ligand specificity of the calcitonin-receptor-like receptor. Nature **393**:333-339.

Melen, K., A. Krogh, and G. von Heijne. 2003. Reliability measures for membrane protein topology prediction algorithms. J Mol Biol **327**:735-744.

Messier, W., and C. B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151-154.

Mi, H., N. Guo, A. Kejariwal, and P. D. Thomas. 2007. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Res **35**:D247-252.

Mi, H., B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano, and P. D. Thomas. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res **33**:D284-288.

Moller, S., M. D. Croning, and R. Apweiler. 2001. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics **17**:646-653.

Moses, A. E., M. R. Wessels, K. Zalcman, S. Alberti, S. Natanson-Yaron, T. Menes, and E. Hanski. 1997. Relative contributions of hyaluronic acid capsule and M protein to virulence in a mucoid strain of the group A Streptococcus. Infect Immun **65**:64-71.

Mullan, L. J., and A. J. Bleasby. 2002. Short EMBOSS User Guide. European Molecular Biology Open Software Suite. Brief Bioinform **3**:92-94.

Nagaraj, S. H., R. B. Gasser, and S. Ranganathan. 2007. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform **8**:6-21.

Ng, P. C., and S. Henikoff. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res **31**:3812-3814.

Nielsen, R. 2006. Comparative genomics: difference of expression. Nature **440**:161.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. S. J, M. D. Adams, and

M. Cargill. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol **3**:e170.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante, and A. G. Clark. 2007. Recent and ongoing selection in the human genome. Nat Rev Genet **8**:857-868.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929-936.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol **302**:205-217.

Ohta, T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol **40**:56-63.

Ohta, T., and J. H. Gillespie. 1996. Development of Neutral and Nearly Neutral Theories. Theor Popul Biol **49**:128-142.

Olson, M. V. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet **64**:18-23.

Olszewski, M. A., J. Gray, and D. J. Vestal. 2006. In silico genomic analysis of the human and murine guanylate-binding protein (GBP) gene clusters. J Interferon Cytokine Res **26**:328-352.

Pearce-Duvet, J. M. 2006. The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. Biol Rev Camb Philos Soc **81**:369-382.

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics **19**:651-652.

Pezet, A., H. Favre, P. A. Kelly, and M. Edery. 1999. Inhibition and restoration of prolactin signal transduction by suppressors of cytokine signaling. J Biol Chem **274**:24497-24502.

Pino, P., I. Vouldoukis, N. Dugas, M. Conti, J. Nitcheu, B. Traore, M. Danis, B. Dugas, and D. Mazier. 2004. Induction of the CD23/nitric oxide pathway in endothelial cells downregulates ICAM-1 expression and decreases cytoadherence of Plasmodium falciparum-infected erythrocytes. Cell Microbiol **6**:839-848.

Prasad, C. K., M. Mahadevan, M. C. MacNicol, and A. M. MacNicol. 2008. Mos 3' UTR regulatory differences underlie species-specific temporal patterns of Mos mRNA cytoplasmic polyadenylation and translational recruitment during oocyte maturation. Mol Reprod Dev **75**:1258-1268.

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. Genetics **160**:1179-1189.

Ramensky, V., P. Bork, and S. Sunyaev. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res **30**:3894-3900.

Rohde, H. M., F. Y. Cheong, G. Konrad, K. Paiha, P. Mayinger, and G. Boehmelt. 2003. The human phosphatidylinositol phosphatase SAC1 interacts with the coatomer I complex. J Biol Chem **278**:52689-52699.

Saar, K., A. Beck, M. T. Bihoreau, E. Birney, D. Brocklebank, Y. Chen, E. Cuppen, S. Demonchy, J. Dopazo, P. Flicek, M. Foglio, A. Fujiyama, I. G. Gut, D. Gauguier, R. Guigo, V. Guryev, M. Heinig, O. Hummel, N. Jahn, S. Klages, V. Kren, M. Kube, H. Kuhl, T. Kuramoto, Y. Kuroki, D. Lechner, Y. A. Lee, N. Lopez-Bigas, G. M. Lathrop, T. Mashimo, I. Medina, R. Mott, G. Patone, J. A. Perrier-Cornet, M. Platzer, M. Pravenec, R. Reinhardt, Y. Sakaki, M. Schilhabel, H. Schulz, T. Serikawa, M. Shikhagaie, S. Tatsumoto, S. Taudien, A. Toyoda, B. Voigt, D. Zelenika, H. Zimdahl, and N. Hubner. 2008. SNP and haplotype mapping for genetic analysis in the rat. Nat Genet **40**:560-566.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature **419**:832-837.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. Positive natural selection in the human lineage. Science **312**:1614-1620.

Sabeti, P. C.P. VariilyB. FryJ. LohmuellerE. HostetterC. CotsapasX. XieE. H. ByrneS. A. McCarrollR. GaudetS. F. SchaffnerE. S. LanderK. A. FrazerD. G. BallingerD. R. CoxD. A. HindsL. L. StuveR. A. GibbsJ. W. BelmontA. BoudreauP. HardenbolS. M. LealS. PasternakD. A. WheelerT. D. WillisF. YuH. YangC. ZengY. GaoH. HuW. HuC. LiW. LinS. LiuH. PanX. TangJ. WangW. WangJ. YuB. ZhangQ. ZhangH. ZhaoJ. ZhouS. B. GabrielR. BarryB. BlumenstielA. CamargoM. DefeliceM. FaggartM. GoyetteS. GuptaJ. MooreH. NguyenR. C. OnofrioM. ParkinJ. RoyE. StahlE. WinchesterL. ZiaugraD. AltshulerY. ShenZ. YaoW. HuangX. ChuY. HeL. JinY. LiuW. SunH. WangY. WangX. XiongL. XuM. M. WayeS. K. TsuiH. XueJ. T. WongL. M. GalverJ. B. FanK. GundersonS. S. MurrayA. R. OliphantM. S. CheeA. MontpetitF. ChagnonV. FerrettiM. LeboeufJ. F. OlivierM. S. PhillipsS. RoumyC. SalleeA. VernerT. J. HudsonP. Y. KwokD. CaiD. C. KoboldtR. D. MillerL. PawlikowskaP. Taillon-MillerM. XiaoL. C. TsuiW. MakY. Q. SongP. K. TamY. NakamuraT. KawaguchiT. KitamotoT. MorizonoA. NagashimaY. OhnishiA. SekineT. TanakaT. TsunodaP. DeloukasC. P. BirdM. DelgadoE. T. DermitzakisR. GwilliamS. HuntJ. MorrisonD. PowellB. E. StrangerP. WhittakerD. R. BentleyM. J. DalyP. I. de BakkerJ. BarrettY. R. ChretienJ. MallerS. McCarrollN. PattersonI. Pe'erA. PriceS. PurcellD. J. RichterP. SabetiR. SaxenaP. C. ShamL. D. SteinL.

KrishnanA. V. SmithM. K. Tello-RuizG. A. ThorissonA. ChakravartiP. E. ChenD. J. CutlerC. S. KashukS. LinG. R. AbecasisW. GuanY. LiH. M. MunroZ. S. QinD. J. ThomasG. McVeanA. AutonL. BottoloN. CardinS. EyheramendyC. FreemanJ. MarchiniS. MyersC. SpencerM. StephensP. DonnellyL. R. CardonG. ClarkeD. M. EvansA. P. MorrisB. S. WeirT. A. JohnsonJ. C. MullikinS. T. SherryM. FeoloA. SkolH. ZhangI. MatsudaY. FukushimaD. R. MacerE. SudaC. N. RotimiC. A. AdebamowoI. AjayiT. AniagwuP. A. MarshallC. NkwodimmahC. D. RoyalM. F. LeppertM. DixonA. PeifferR. QiuA. KentK. KatoN. NiikawaI. F. AdewoleB. M. KnoppersM. W. FosterE. W. ClaytonJ. WatkinD. MuznyL. NazarethE. SodergrenG. M. WeinstockI. YakubB. W. BirrenR. K. WilsonL. L. FultonJ. RogersJ. BurtonN. P. CarterC. M. CleeM. GriffithsM. C. JonesK. McLayR. W. PlumbM. T. RossS. K. SimsD. L. WilleyZ. ChenH. HanL. KangM. GodboutJ. C. WallenburgP. L'ArchevequeG. BellemareK. SaekiD. AnH. FuQ. LiZ. WangR. WangA. L. HoldenL. D. BrooksJ. E. McEwenM. S. GuyerV. O. WangJ. L. PetersonM. ShiJ. SpiegelL. M. SungL. F. ZachariaF. S. CollinsK. KennedyR. Jamieson, and J. Stewart. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature **449**:913-918.

Sabeti, P. C., E. Walsh, S. F. Schaffner, P. Varilly, B. Fry, H. B. Hutcheson, M. Cullen, T. S. Mikkelsen, J. Roy, N. Patterson, R. Cooper, D. Reich, D. Altshuler, S. O'Brien, and E. S. Lander. 2005. The case for selection at CCR5-Delta32. PLoS Biol **3**:e378.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**:406-425.

Saleh, M., J. P. Vaillancourt, R. K. Graham, M. Huyck, S. M. Srinivasula, E. S. Alnemri, M. H. Steinberg, V. Nolan, C. T. Baldwin, R. S. Hotchkiss, T. G. Buchman, B. A. Zehnbauer, M. R. Hayden, L. A. Farrer, S. Roy, and D. W. Nicholson. 2004. Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. Nature **429**:75-79.

Scrimshaw, N. S., and E. B. Murray. 1988. The acceptability of milk and milk products in populations with a high prevalence of lactose intolerance. Am J Clin Nutr **48**:1079-1159.

Shriver, M. D., G. C. Kennedy, E. J. Parra, H. A. Lawson, V. Sonpar, J. Huang, J. M. Akey, and K. W. Jones. 2004. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics **1**:274-286.

Slatkin, M., and G. Bertorelle. 2001. The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics **158**:865-874.

Sommer, S., C. B. Pudrith, C. J. Colvin, and P. M. Coussens. 2009. Mycobacterium avium subspecies paratuberculosis suppresses expression of IL-12p40 and iNOS genes induced by signalling through CD40 in bovine monocyte-derived macrophages. Vet Immunol Immunopathol **128**:44-52.

Soranzo, N., F. Rivadeneira, U. Chinappen-Horsley, I. Malkina, J. B. Richards, N. Hammond, L. Stolk, A. Nica, M. Inouye, A. Hofman, J. Stephens, E. Wheeler, P. Arp, R. Gwilliam, P. M. Jhamai, S. Potter, A. Chaney, M. J. Ghori, R. Ravindrarajah, S. Ermakov, K. Estrada, H. A. Pols, F. M. Williams, W. L. McArdle, J. B. van Meurs, R. J. Loos, E. T. Dermitzakis, K. R. Ahmadi, D. J. Hart, W. H. Ouwehand, N. J. Wareham, I. Barroso, M. S. Sandhu, D. P. Strachan, G. Livshits, T. D. Spector, A. G. Uitterlinden, and P. Deloukas. 2009. Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. PLoS Genet **5**:e1000445.

Sporri, B., P. E. Kovanen, A. Sasaki, A. Yoshimura, and W. J. Leonard. 2001. JAB/SOCS1/SSI-1 is an interleukin-2-induced inhibitor of IL-2 signaling. Blood **97**:221-226.

Stauffer, R. L., A. Walker, O. A. Ryder, M. Lyons-Weiler, and S. B. Hedges. 2001. Human and ape molecular clocks and constraints on paleontological hypotheses. J Hered **92**:469-474.

Stone, E. A., and A. Sidow. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res **15**:978-986.

Stone, R. T., W. M. Grosse, E. Casas, T. P. Smith, J. W. Keele, and G. L. Bennett. 2002. Use of bovine EST data and human genomic sequences to map 100 gene-specific bovine markers. Mamm Genome **13**:211-215.

Suzuki, Y., and T. Gojobori. 1999. A method for detecting positive selection at single amino acid sites. Mol Biol Evol **16**:1315-1328.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585-595.

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol **24**:1596-1599.

Tate, G., T. Suzuki, K. Kishimoto, and T. Mitsuya. 2004. Novel mutations of EVER1/TMC6 gene in a Japanese patient with epidermodysplasia verruciformis. J Hum Genet **49**:223-225.

Thomas, P. D., M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. 2003a. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res **13**:2129-2141.

Thomas, P. D., A. Kejariwal, M. J. Campbell, H. Mi, K. Diemer, N. Guo, I. Ladunga, B. Ulitsky-Lazareva, A. Muruganujan, S. Rabkin, J. A. Vandergriff, and O. Doremieux. 2003b. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res **31**:334-341.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**:4673-4680.

Tournamille, C., Y. Colin, J. P. Cartron, and C. Le Van Kim. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet **10**:224-228.

Troy, C. S., D. E. MacHugh, J. F. Bailey, D. A. Magee, R. T. Loftus, P. Cunningham, A. T. Chamberlain, B. C. Sykes, and D. G. Bradley. 2001. Genetic evidence for Near-Eastern origins of European cattle. Nature **410**:1088-1091.

van Spriel, A. B., K. L. Puls, M. Sofi, D. Pouniotis, H. Hochrein, Z. Orinska, K. P. Knobeloch, M. Plebanski, and M. D. Wright. 2004. A regulatory role for CD37 in T cell proliferation. J Immunol **172**:2953-2961.

Van Tassell, C. P., T. P. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren, and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods **5**:247-252.

Venter, J. C.M. D. AdamsE. W. MyersP. W. LiR. J. MuralG. G. SuttonH. O. SmithM. YandellC. A. EvansR. A. HoltJ. D. GocayneP. AmanatidesR. M. BallewD. H. HusonJ. R. WortmanQ. ZhangC. D. KodiraX. H. ZhengL. ChenM. SkupskiG. SubramanianP. D. ThomasJ. ZhangG. L. Gabor MiklosC. NelsonS. BroderA. G. ClarkJ. NadeauV. A. McKusickN. ZinderA. J. LevineR. J. RobertsM. SimonC. SlaymanM. HunkapillerR. BolanosA. DelcherI. DewD. FasuloM. FlaniganL. FloreaA. HalpernS. HannenhalliS. KravitzS. LevyC. MobarryK. ReinertK. RemingtonJ. Abu-ThreidehE. BeasleyK. BiddickV. BonazziR. BrandonM. CargillI. ChandramouliswaranR. CharlabK. ChaturvediZ. DengV. Di FrancescoP. DunnK. EilbeckC. EvangelistaA. E. GabrielianW. GanW. GeF. GongZ. GuP. GuanT. J. HeimanM. E. HigginsR. R. JiZ. KeK. A. KetchumZ. LaiY. LeiZ. LiJ. LiY. LiangX. LinF. LuG. V. MerkulovN. MilshinaH. M. MooreA. K. NaikV. A. NarayanB. NeelamD. NusskernD. B. RuschS. SalzbergW. ShaoB. ShueJ. SunZ. WangA. WangX. WangJ. WangM. WeiR. WidesC. XiaoC. YanA. YaoJ. YeM. ZhanW. ZhangH. ZhangQ. ZhaoL. ZhengF. ZhongW. ZhongS. ZhuS. ZhaoD. GilbertS. BaumhueterG. SpierC. CarterA. CravchikT. WoodageF. AliH. AnA. AweD. BaldwinH. BadenM. BarnsteadI. BarrowK. BeesonD. BusamA. CarverA. CenterM. L. ChengL. CurryS. DanaherL. DavenportR. DesiletsS. DietzK. DodsonL. DoupS. FerrieraN. GargA. GluecksmannB. HartJ. HaynesC. HaynesC. HeinerS. HladunD. HostinJ. HouckT. HowlandC. IbegwamJ. JohnsonF. KalushL. KlineS. KoduruA. LoveF. MannD. MayS. McCawleyT. McIntoshI. McMullenM. MoyL. MoyB. MurphyK. NelsonC. PfannkochE. PrattsV. PuriH. QureshiM. ReardonR. RodriguezY. H. RogersD. RombladB. RuhfelR. ScottC. SitterM. SmallwoodE. StewartR. StrongE. SuhR. ThomasN. N. TintS. TseC. VechG. WangJ. WetterS. WilliamsM. WilliamsS. WindsorE. Winn-DeenK. WolfeJ. ZaveriK. ZaveriJ. F. AbrilR. GuigoM. J. CampbellK. V. SjolanderB. KarlakA. KejariwalH. MiB. LazarevaT. HattonA.

NarechaniaK. DiemerA. MuruganujanN. GuoS. SatoV. BafnaS. IstrailR. LippertR. SchwartzB. WalenzS. YoosephD. AllenA. BasuJ. BaxendaleL. BlickM. CaminhaJ. Carnes-StineP. CaulkY. H. ChiangM. CoyneC. DahlkeA. MaysM. DombroskiM. DonnellyD. ElyS. EsparhamC. FoslerH. GireS. GlanowskiK. GlasserA. GlodekM. GorokhovK. GrahamB. GropmanM. HarrisJ. HeilS. HendersonJ. HooverD. JenningsC. JordanJ. JordanJ. KashaL. KaganC. KraftA. LevitskyM. LewisX. LiuJ. LopezD. MaW. MajorosJ. McDanielS. MurphyM. NewmanT. NguyenN. NguyenM. NodellS. PanJ. PeckM. PetersonW. RoweR. SandersJ. ScottM. SimpsonT. SmithA. SpragueT. StockwellR. TurnerE. VenterM. WangM. WenD. WuM. WuA. XiaA. Zandieh, and X. Zhu. 2001. The sequence of the human genome. Science **291**:1304-1351.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol **4**:e72.

Wang, P. J., J. R. McCarrey, F. Yang, and D. C. Page. 2001. An abundance of X-linked genes expressed in spermatogonia. Nat Genet **27**:422-426.

Waterston, R. H.K. Lindblad-TohE. BirneyJ. RogersJ. F. AbrilP. AgarwalR. AgarwalaR. AinscoughM. AlexanderssonP. AnS. E. AntonarakisJ. AttwoodR. BaertschJ. BaileyK. BarlowS. BeckE. BerryB. BirrenT. BloomP. BorkM. BotcherbyN. BrayM. R. BrentD. G. BrownS. D. BrownC. BultJ. BurtonJ. ButlerR. D. CampbellP. CarninciS. CawleyF. ChiaromonteA. T. ChinwallaD. M. ChurchM. ClampC. CleeF. S. CollinsL. L. CookR. R. CopleyA. CoulsonO. CouronneJ. CuffV. CurwenT. CuttsM. DalyR. DavidJ. DaviesK. D. DelehauntyJ. DeriE. T. DermitzakisC. DeweyN. J. DickensM. DiekhansS. DodgeI. DubchakD. M. DunnS. R. EddyL. ElnitskiR. D. EmesP. EswaraE. EyrasA. FelsenfeldG. A. FewellP. FlicekK. FoleyW. N. FrankelL. A. FultonR. S. FultonT. S. FureyD. GageR. A. GibbsG. GlusmanS. GnerreN. GoldmanL. GoodstadtD. GrafhamT. A. GravesE. D. GreenS. GregoryR. GuigoM. GuyerR. C. HardisonD. HausslerY. HayashizakiL. W. HillierA. HinrichsW. HlavinaT. HolzerF. HsuA. HuaT. HubbardA. HuntI. JacksonD. B. JaffeL. S. JohnsonM. JonesT. A. JonesA. JoyM. KamalE. K. KarlssonD. KarolchikA. KasprzykJ. KawaiE. KeiblerC. KellsW. J. KentA. KirbyD. L. KolbeI. KorfR. S. KucherlapatiE. J. KulbokasD. KulpT. LandersJ. P. LegerS. LeonardI. LetunicR. LevineJ. LiM. LiC. LloydS. LucasB. MaD. R. MaglottE. R. MardisL. MatthewsE. MauceliJ. H. MayerM. McCarthyW. R. McCombieS. McLarenK. McLayJ. D. McPhersonJ. MeldrimB. MeredithJ. P. MesirovW. MillerT. L. MinerE. MonginK. T. MontgomeryM. MorganR. MottJ. C. MullikinD. M. MuznyW. E. NashJ. O. NelsonM. N. NhanR. NicolZ. NingC. NusbaumM. J. O'ConnorY. OkazakiK. OliverE. Overton-LartyL. PachterG. ParraK. H. PepinJ. PetersonP. PevznerR. PlumbC. S. PohlA. PoliakovT. C. PonceC. P. PontingS. PotterM. QuailA. ReymondB. A. RoeK. M. RoskinE. M. RubinA. G. RustR. SantosV. SapojnikovB. SchultzJ. SchultzM. S. SchwartzS. SchwartzC. ScottS. SeamanS. SearleT. SharpeA. SheridanR. ShownkeenS. SimsJ. B. SingerG. SlaterA. SmitD. R. SmithB. SpencerA. StabenauN. Stange-ThomannC. SugnetM. SuyamaG. TeslerJ. ThompsonD. TorrentsE. TrevaskisJ. TrompC. UclaA. Ureta-VidalJ. P. VinsonA. C. Von NiederhausernC. M. WadeM. WallR. J. WeberR. B. WeissM. C. WendlA. P.

WestK. WetterstrandR. WheelerS. WhelanJ. WierzbowskiD. WilleyS. WilliamsR. K. WilsonE. WinterK. C. WorleyD. WymanS. YangS. P. YangE. M. ZdobnovM. C. Zody, and E. S. Lander. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

Wayne, C. M., J. A. MacLean, G. Cornwall, and M. F. Wilkinson. 2002. Two novel human X-linked homeobox genes, hPEPP1 and hPEPP2, selectively expressed in the testis. Gene **301**:1-11.

Werner, F. A., G. Durstewitz, F. A. Habermann, G. Thaller, W. Kramer, S. Kollers, J. Buitkamp, M. Georges, G. Brem, J. Mosner, and R. Fries. 2004. Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. Anim Genet **35**:44-49.

Wessels, M. R., A. E. Moses, J. B. Goldberg, and T. J. DiCesare. 1991. Hyaluronic acid capsule is a virulence factor for mucoid group A streptococci. Proc Natl Acad Sci U S A **88**:8317-8321.

Williams, T. M., and M. P. Lisanti. 2004. The caveolin proteins. Genome Biol **5**:214.

Wong, G. K.B. LiuJ. WangY. ZhangX. YangZ. ZhangQ. MengJ. ZhouD. LiJ. ZhangP. NiS. LiL. RanH. LiR. LiH. ZhengW. LinG. LiX. WangW. ZhaoJ. LiC. YeM. DaiJ. RuanY. ZhouY. LiX. HeX. HuangW. TongJ. ChenJ. YeC. ChenN. WeiL. DongF. LanY. SunZ. YangY. YuY. HuangD. HeY. XiD. WeiQ. QiW. LiJ. ShiM. WangF. XieX. ZhangP. WangY. ZhaoN. LiN. YangW. DongS. HuC. ZengW. ZhengB. HaoL. W. HillierS. P. YangW. C. WarrenR. K. WilsonM. BrandstromH. EllegrenR. P. CrooijmansJ. J. van der PoelH. BovenhuisM. A. GroenenI. OvcharenkoL. GordonL. StubbsS. LucasT. GlavinaA. AertsP. KaiserL. RothwellJ. R. YoungS. RogersB. A. WalkerA. van HaterenJ. KaufmanN. BumsteadS. J. LamontH. ZhouP. M. HockingD. MorriceD. J. de KoningA. LawN. BartleyD. W. BurtH. HuntH. H. ChengU. GunnarssonP. WahlbergL. AnderssonE. KindlundM. T. TammiB. AnderssonC. WebberC. P. PontingI. M. OvertonP. E. BoardmanH. TangS. J. HubbardS. A. WilsonJ. Yu, and H. Yang. 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature **432**:717-722.

Woolfit, M., and L. Bromham. 2005. Population size and molecular evolution on islands. Proc Biol Sci **272**:2277-2282.

Wright, S. 1951. The genetical structure of populations. Ann. Eugen. **15**:323-354.

Xu, E. W., P. Kearney, and D. G. Brown. 2006. The use of functional domains to improve transmembrane protein topology prediction. J Bioinform Comput Biol **4**:109-123.

Xue, Y., A. Daly, B. Yngvadottir, M. Liu, G. Coop, Y. Kim, P. Sabeti, Y. Chen, J. Stalker, E. Huckle, J. Burton, S. Leonard, J. Rogers, and C. Tyler-Smith. 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet **78**:659-670.

Yang, S., Y. Liu, A. A. Lin, L. L. Cavalli-Sforza, Z. Zhao, and B. Su. 2005. Adaptive evolution of MRGX2, a human sensory neuron specific gene involved in nociception. Gene **352**:30-35.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**:555-556.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**:1586-1591.

Yang, Z. 2002. Inference of selection from multiple species alignments. Curr Opin Genet Dev **12**:688-694.

Yang, Z., and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol **19**:908-917.

Yen, C. L., S. J. Stone, S. Koliwad, C. Harris, and R. V. Farese, Jr. 2008. Thematic review series: glycerolipids. DGAT enzymes and triacylglycerol biosynthesis. J Lipid Res **49**:2283-2301.

Yunis, J. J., and O. Prakash. 1982. The origin of man: a chromosomal pictorial legacy. Science **215**:1525-1530.

Yunis, J. J., J. R. Sawyer, and K. Dunham. 1980. The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. Science **208**:1145-1148.

Zhang, J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol **21**:1332-1339.

Zhang, J., S. Kumar, and M. Nei. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol Biol Evol **14**:1335-1338.

Zhang, J., and M. Nei. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol **44 Suppl 1**:S139-146.

Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol **22**:2472-2479.

Zhang, X., M. Cairns, B. Rose, C. O'Brien, K. Shannon, J. Clark, J. Gamble, and N. Tran. 2009. Alterations in miRNA processing and expression in pleomorphic adenomas of the salivary gland. Int J Cancer **124**:2855-2863.

Zhou, J., T. Zhu, C. Hu, H. Li, G. Chen, G. Xu, S. Wang, and D. Ma. 2008. Comparative genomics and function analysis on BI1 family. Comput Biol Chem **32**:159-162.

List of Supplementary Tables and Documentation

This is a list of all files contained in the accompanying CD to the thesis, including supplementary documentation and supplementary tables (STs).

Supp. Doc.	SOM for both Bovine HapMap and Bovine Genome projects
ST 2.1	EC domain PAML analysis of free-ratios model (m1) vs null model (m0)
ST 2.2	EC domain human and chimpanzee pair-wise PAML analysis
ST 2.3	EC domain PAML analysis of free-ratios model (m1) vs null model (m0) for human, chimpanzee and macaque
ST 2.4	EC domain GO Over-Representation Analysis (ORA)
ST 3.1	HMRDC EC domain GO ORA
ST 3.2	BGP variable selection
ST 3.3	BGP d_N/d_S
ST 3.4	BGP bovine lineage model (m2) vs null model (m0)
ST 3.5	BGP free-ratios model (m1) vs null model (m0)
ST 3.6	BGP GO biological processes MWU
ST 3.7	BGP GO molecular functions MWU
ST 3.8	BGP PANTHER MWU
ST 3.9	BGP InnateDB pathways
ST 4.1	Bovine EST InnateDB Ontology ORA
ST 5.1	33K LSBL Top 1% (each population per tab)
ST 5.2	33K LSBL Top 1% GO ORA for African Population
ST 5.3	33K LSBL Top 1% GO ORA for European Population
ST 5.4	33K LSBL Top 1% GO ORA for Indicine Population
ST 5.5	50K LSBL, Fay & Wu's *H*, composite calculation for African population
ST 5.6	50K LSBL, Fay & Wu's *H*, composite calculation for European population
ST 5.7	50K LSBL, Fay & Wu's *H*, composite calculation for Indian population
ST 5.8	50K Top 1% GO ORA for African Population
ST 5.9	50K Top 1% GO ORA for European Population
ST 5.10	50K Top 1% GO ORA for African Population