Different patterns of gain and loss in genomic evolution

by

David González Knowles

A thesis submitted to
The University of Dublin
for the degree of

Doctor of Philosophy

Smurfit Institute of Genetics
Trinity College
University of Dublin

April, 2008

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and David González Knowles.

Signature of Author...

David González Knowles
30 April, 2008

Summary

Evolutionary forces may act on the genome at different levels, from the change of single nucleotides to the duplication or rearrangement of whole chromosomes. Some of these evolutionary processes can be responsible for changes at different levels, while others will be more specific.

In this work we will examine changes in the genome at two different levels, at the genomic level by identifying changes in gene order and content, and at the gene level by identifying cases of intron gain and loss. Identification of these changes as well as the mechanisms that are responsible for them will allow a better understanding of the processes involved in the evolution of the genome.

In the first part of this work we used the genome sequence available for the three primates *Homo sapiens*, *Pan troglodytes* and *Macaca mulatta* in order to build a set of synteny blocks where gene order is conserved. This was done for each pair of species and the resulting blocks confirmed the high level of similarity between the three genomes, with most of their genes included within the blocks.

We classified all protein-coding genes within the synteny blocks and identified those that are conserved between each pair of species as well as those that are present on the three species as one-to-one orthologues, which included more than half of the human genes. We also identified gene duplications and translocation within the identified synteny blocks. Where differences in gene content were found we examined the assembly information and genome annotation in order to determine which of these differences were reliable and which ones may be due to assembly or annotation errors.

In the case of human and chimpanzee, those cases that were identified as reliable differences between the two species were compared to the macaque in order to identify lineage specific differences that had occurred since the chimpanzee and human lineages diverged. These lineage specific genes were further examined and we identified nine cases in which an origin by exaptation of non-coding DNA could not be ruled out.

We also searched for cases of alternatively spliced genes that may have become duplicated and undergone subsequent subfunctionalization, by differential loss of splice variants, since the divergence of the human and chimpanzee lineage. We identified one alternatively spliced gene that was duplicated specifically in the human lineage where the two copies show different alternative splice forms annotated. However, we could not find unambiguous evidence at the sequence level of the inability of either copy to produce all of these variants.

In the second part of this study we identified those introns that had been differentially gained or lost between pairs of paralogous genes that originated simultaneously in a large genome duplication that occurred in *Arabidopsis thaliana* 20 - 60 Mya.

We found a high rate of intron turnover since the duplication event, although with the available data we were only able to identify a small fraction of the inserted/deleted introns unambiguously as gains our losses. Despite the relatively recent origin of the new introns we identified, only in one case were we able to identify the precise mechanism by which a new intron had originated.

Acknowledgements

I would like to thank Aoife for her supervision of the project, as well as all the members of the Molecular Evolution Lab, particularly Kirsten for all the proofreading, Åsa, Takashi and Daniel.

I would also like to thank all my friends from Spain for their support in my decision to come to Ireland as well as all the great friends I have met while I was here in Ireland for their patience and help.

Also I would like to thank the current and former members of the Trinity College Karate club for the great opportunity the training sessions offered for reducing the stress levels as well as clearing the mind.

Finally thanks to all the people attending the Wednesday seminars for their comments and suggestions. In particular Ken Wolfe for some very helpful ideas.

And of course I would like to thank my family both for their support of my decision to come to Ireland and for their support during all the time I have spent here.

Contents

List of Tables

List of Figures

Abbreviations

aa	Amino acid
BAC	Bacterial artificial chromosome
BLAST	Basic local alignment tool
BLAT	BLAST like alignment tool
CCL	Chronic lymphocytic leukemia
cds	Coding DNA sequence
CG	Chorionic gonadotropin
DNA	Deoxyribonucleic acid
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
EST	Expressed sequence tag
GDH	Glutamate dehydrogenase
GEO	Gene expression omnibus
HG	Human genome
IL	Interleukin
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
LUCA	Last universal common ancestor
LH	Luteinizing hormone
miRNA	Micro RNA
mRNA	Messenger RNA
Mya	Million years ago

Myr	Million years
NMD	Nonsense mediated decay
indel	Insertion/deletion
IO	Initial orthologue
ORF	Open reading frame
pg	Picograms
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
RTH	Reciprocal top hit
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
TAIR	The Arabidiopsis Information Resource
TE	Transposable element
tRNA	transfer RNA
UTR	Un-translated region
WD	Working draft
WGS	Whole genome shotgun

Every day you may make progress.

Every step may be fruitful.

Yet there will stretch out before you an ever-lengthening, ever-ascending, ever-improving path.

You know you will never get to the end of the journey.

But this, so far from discouraging, only adds to the joy and glory of the climb.

Sir Winston Churchill

Chapter 1

Introduction

1.1 Historical perspective of molecular evolution

The mechanism by which the genome of a living organism evolves has drawn a lot of interest since the nature of the hereditary material was discovered and the genetic code deciphered during the 1950s. However the more we learn about the genome the more complex it appears, from the initial discovery that DNA content did not correlate with organism complexity that gave rise to the "C-value paradox" to the more recent discoveries of the implication of small non-coding RNAs in gene regulation, we have advanced a long way. The "C-value paradox" was resolved with the discovery of non-coding DNA. The complete sequencing of the human genome (Lander *et al.*, 2001), another a milestone in modern molecular biology, gave birth to a new era of comparative genomics, and a huge growth in the amount of publicly available sequence data. However, as so often happens with important scientific achievements, many new questions have been raised with our increased understanding of the genome that still await a satisfactory answer. The lack of understanding of how the genome really works being clearly evidenced by the difficulty of obtaining, even several years after the official completion of the human sequence, an unambiguous value for the number of genes present in the genome (Guigo *et al.*, 2006).

Notwithstanding the absence of a complete understanding of genome intricacies, the discovery of the nature of the hereditary material during the last century provided a tangible explanation for the transfer of information from one generation to the next. This was one of the main elements lacking in the theory of evolution proposed on 1858 by Charles Darwin in his work *The Origin of Species*.

Darwin's theory of evolution states that every organism originates from another organism by means of "descent with modification". This modification is now known to be caused by changes at the genomic level. It follows from this that the hereditary material is a historical record that runs back uninterrupted from any of the extant species in the planet to the origin of all cellular life. The recent advances in molecular genetics which have provided us with the means to determine the sequence of DNA molecules as well as proteins give us the means to read this historical record. This allows us to study not only the evolution of specific molecules, but also the reconstruction of the evolutionary history of genes and complete organisms. This is not as easy as it may sound, because the historical record genetic material presents us with has been compiled by writing all those changes that have occurred over previous records. Although the most recent changes are readily identifiable the further back in time we look it becomes increasingly difficult to retrieve meaningful information about the changes that have occurred. However it still provides overwhelming support for the whole field of evolutionary studies, which even up to the 1970s were still considered in some scientific circles as unwarranted assumptions and speculations (Graur and Li, 2000).

Regardless of the difficulties that exist when using genomic information for the study of events that have taken place far back in time, it allows us to achieve a huge increase in both depth and detail when searching for conserved and divergent features between different species when we compare it to the previously available classification methods that were entirely based on phenotypic characters. This becomes very clear when observing species with different degrees of relatedness within a large group such as the Eukaryotes. This group originated from an ancestral Eukaryote that lived approximately 3000 Mya (Nei *et al.*, 2001) and so all Eukaryotes descend from the same original genome. When looking at species within a closely related group such as mammals which diverged 310 Mya (Hedges *et al.*, 2004), certain common traits are readily identifiable phenotypically, and it does not require a great deal of imagination to find common characters shared between all of them that point towards their common ancestry. However this is much more difficult when we try to find similarities between two distant Eukaryote groups such as plants (Plantae) and animals (Animalia) that diverged around 1300 - 1600 Mya (Nei *et al.* 2001, Hedges *et al.* 2004). If we try to use phenotypic characters, we are unlikely to find any meaningful ones. However if we look deeper into the cellular structure we can find some similarities, and by going even deeper into the nucleotide sequences of the different genes we find a surprisingly large number of similarities that are sufficient to establish a meaningful, even if not completely resolved, relation between these distant groups.

When comparing these two distant groups, we should bear in mind that the huge differences between them have been caused by a gradual accumulation of different changes along two different lineages which originally shared the same genomic sequence. These same mechanisms that over 1600 My caused the divergence of plants and animals are the ones responsible for the difference between more recently diverged species.

By comparing genome sequences of extant organisms that are separated by small phylogenetic distances we should be able to determine with a high resolution the changes that have occurred between them. And not only can we identify these differences but also, by using a third species that diverged earlier, the particular lineage in which each of these changes originated.

Although part of the different phenotypic characters in extant species may be caused by the environment in which they develop, most of these differences are a consequence of the different evolutionary paths their genomes have taken since they diverged from their common ancestor (Lewontin, 2000). In order to determine how these genomic differences have been attained we need to know in what ways the genome can change over time. Changes may occur in gene content, genome size, distribution of genes along the genome, regulation of different genes or any combination of these. Each of these changes will occur with a different frequency and cause different effects which may produce visible phenotypic changes or not. However ultimately when sufficient differences have accumulated, the consequence is reproductive isolation and speciation (Lynch and Conery 2000, Taylor *et al.* 2001).

The discovery that amino acid sequences accumulate changes at an approximately constant rate over time allows us to use this genomic record not only to determine what changes occurred in each lineage, but also to estimate the times at which they occurred (Zuckerkandl and Pauling, 1965). This rate, which has been equated to a molecular clock, varies between genes and between species (Li, 1993), which allows for comparisons between closer or more distant species by choosing faster or slower evolving genes. In order to relate this molecular clock to astronomical time it requires calibration, which is usually done by reference to the fossil record. However the accuracy of these estimates decreases with the time since the divergence of the two species, because the evolutionary pressures on a certain gene are likely to change over time, slowing or accelerating this clock. Nonetheless we can obtain rough estimates of actual divergence times which are very useful when no reliable fossil record is available (Nei *et al.*, 2001), and also the relative times of divergences of different species.

The difficulties that arise when estimating exact lineage divergence times increase with the time since the two species diverged as can be seen in the different time estimates obtained for the same lineage splits by different groups using different methods (Nei *et al.* 2001, Hedges *et al.* 2004, Cavalier-Smith 2006).

The availability of the genomic sequence of three primates, macaque (*Macaca mulatta*), chimpanzee (*Pan troglodytes*) and human (*Homo sapiens*) gives us the opportunity to study the changes that have occurred since the divergence of our closest living relative, the chimpanzee and by using macaque as an outgroup (Waddell *et al.* 2001, Reyes *et al.* 2004) determine in which lineage each of these changes occurred.

1.2 Classification of primates

Primates are a diverse group of mammals that diverged from a common ancestor around 60 Mya (Purvis 1995, Hayasaka *et al.* 1988). Currently there are around 220 living species which are divided into two large groups Strepsirrhini ("wet nose"), also called prosimians and Haplorrhini ("simple nose'). The group of the tarsiers (*Tarsius*) is included in the Haplorrhini according to the currently accepted classification, although there has been much debate over the phylogenetic placing of this group, which incidentally has a dry nose (Hayasaka *et al.* 1988, Purvis 1995, Zietkiewicz *et al.* 1999). Haplorrhini are further divided between the New World monkeys, Platyrrhini ("flat-nose") and the Catarrhini ("downward nose"), which is the group on which we will focus (figure 1.1 overleaf).

Catarrhini diverged between 23 and 30 Mya (Hayasaka *et al.* 1988, Purvis 1995, Glazko and Nei 2003, Raaum *et al.* 2005, Gibbs *et al.* 2007) into the Hominoidea and Cercopithecoidea (Old World monkeys). The later group includes the rhesus macaque (*M. mulatta*) which has been frequently used as an animal model for biomedical research and is one of the best studied non-human primates. Hominoidea (apes) includes the gibbons and the great apes (Hominidae). Both chimpanzee, (*P. troglodytes*) and human (*H. sapiens*) which diverged approximately 6 -7 Mya (Purvis 1995, Glazko and Nei 2003, Raaum *et al.* 2005, Gibbs *et al.* 2007) are included together with the gorilla (*Gorilla*) and orangutan (*Pongo*) in this group. Figure 1.1 shows the approximate times of divergence of the main primate groups from the lineage leading to human.

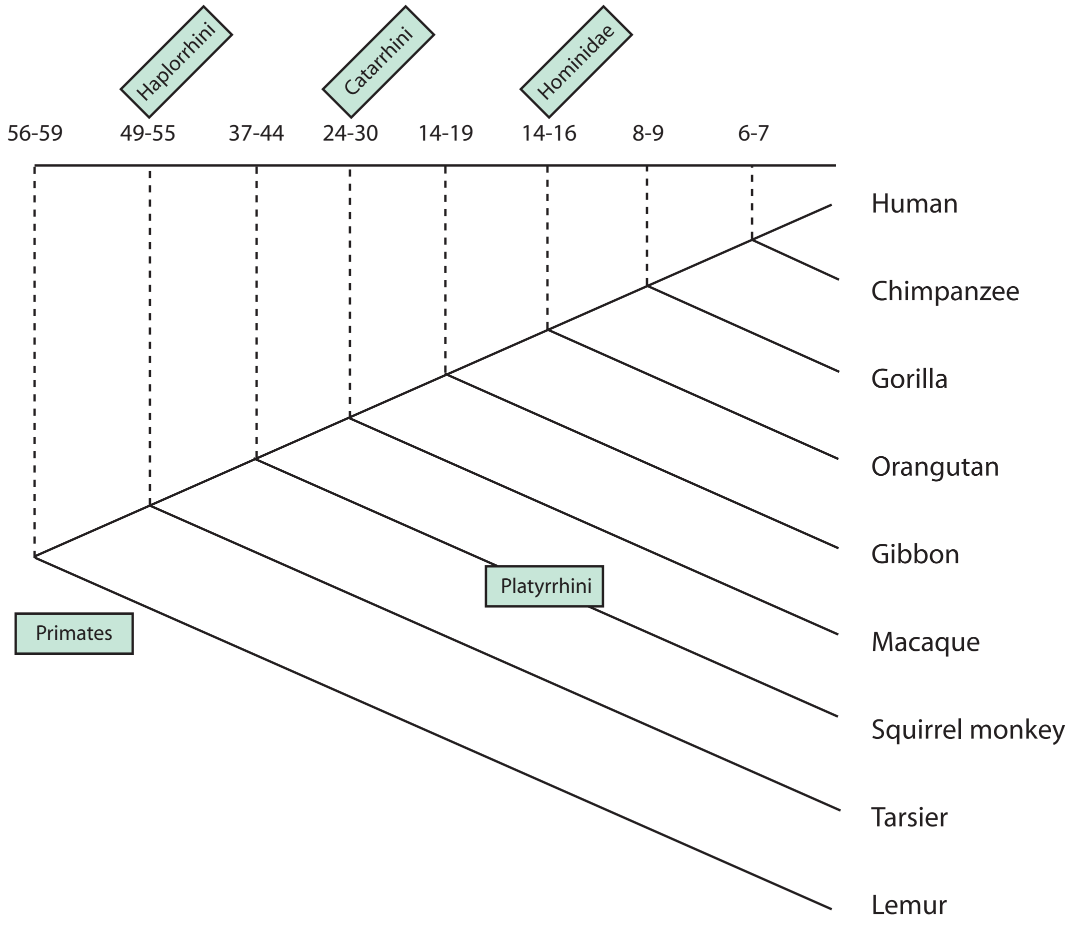


Figure 1.1: Relationship and approximate times of divergence in My before present of the main primate groups. Adapted from Purvis (1995) and Raaum *et al.* (2005)

A draft genomic sequence of the macaque became available on 2007 (Gibbs *et al.*, 2007). The complete human genome sequence has been available for several years (Lander *et al.*, 2001) and a draft genome sequence for the chimpanzee was released in 2005 (CSAC, 2005).

1.3 Genome evolution

The genome, being a blueprint of the organism to which it belongs, does not remain static. Every species evolves in order to better adapt to a continuously changing environment, and the phenotypic changes that occur during this process will reflect the changes that have taken place within its genome. These changes, as noted above, may affect size, gene content, sequence, structure, or any combination of these. However not all changes that occur will become established, for instance changes that occur in somatic cells, regardless of any benefits they may confer, will be lost. In order for any of these changes to become established, they need to occur within the germ-line of the organism, and also produce viable gametes that will allow the generation of fertile offspring. This requirement is likely to filter out the most dramatic changes that may occur. Although it does not mean major changes will never occur, as can be seen from the abundant cases of ancestral polyploidization events that are mentioned in the literature in plants animals and fungi, which are in many cases associated to rapid speciation of the affected lineages (Skrabanek and Wolfe 1998, Blanc *et al.* 2003, Seoighe 2003, Vandepoele *et al.* 2004, Storchova *et al.* 2006, Semon and Wolfe 2007 and others).

1.3.1 Genome size

It has been noted since shortly after the DNA was identified as the hereditary material, that the amount of DNA in eukaryotic cells is not correlated to either the size or the number of different genes within the organism (Mirsky and Ris

1951, Oliver *et al.* 2007). However there are many phenotypic features that are directly correlated to genome size, such as cell volume, or karyoplasmic ration (Cavalier-Smith, 2005). Selection upon these phenotypic characters could maintain the genome of a certain species within certain size regardless of the actual content of its genome (Hughes and Hughes 1995, Cavalier-Smith 2005, Knibbe *et al.* 2007).

Different genome sizes in different species can have various causes, such as differences in chromosome number, ploidy (number of sets of chromosomes), number of genes, or fraction of non-coding DNA. Each of these features may confer certain advantages and certain disadvantages to the organism.

Genome size may be altered by indels transposable elements or duplications. Some of these will change the gene number and some will not. Most of these mechanisms cause an increase in the genome size (Imai *et al.*, 2007), whilst a few such as unequal recombination have been linked to a decrease in genome size (Devos *et al.*, 2002). The distribution of non-coding DNA is different in plants from that in animals, where a much higher fraction of the non-coding DNA belongs to intronic regions (McLysaght *et al.* 2000, Wong *et al.* 2000, Wendel *et al.* 2002), which could indicate a difference in the main contributors to variations in the genome size. It could also be subject to different selective pressures in unicellular versus multicellular organisms (Lynch, 2005).

The range of genome sizes present in mammals varies from 1.7 pg. to 8.4 with an average of 3.5 –which is the size of the human genome. The chimpanzee genome is slightly larger on average and the macaque genome slightly smaller (Gregory, T.R. 2008. Animal Genome Size Database. `http://www.genomesize.com`). No significant intra-specific variation in DNA content has been observed in primates, with the exception of the small difference between males and females caused by the difference in sizes between the X and Y chromosomes.

1.3.2 Gene content

Gene content is another important factor in the evolution of the genome. Protein-coding gene content varies between different species, and the importance of changes in the protein repertoire as a major cause of phenotypic differences and evolutionary change has been recognized for a long time.

The similarities and differences in the protein-coding gene content between two species can be evaluated using different approaches. The number of orthologues that are shared between the two species can be compared (Snel *et al.*, 1999), and this is probably the most frequently used method. Another approach is the comparison of the gene families present and absent in the different genomes. This can also be informative (Hughes and Friedman, 2004), and may be more useful in the case of incomplete gene sets, as even if individual genes are missing it will be less likely for large gene families to be completely absent.

However any method that intends to determine the differences between two different organisms will always rely on the annotation of the genes in each of the species' genomes, and unless this annotation is correct any results obtained will be unreliable. The available annotation is far from perfect as has been shown in several studies (Guigo *et al.*, 2006) and there may be an overestimation of the number of protein-coding sequences in some vertebrate lineages due to the annotation of genes as protein-coding when they are are transcribed but not translated (Goodstadt and Ponting, 2006).

Gene content within the genome is not static, and the number of genes present may increase or decrease over time. Several different mechanisms have been suggested that can generate new genes, and these are summarized below and reviewed by (Long *et al.*, 2003).

- **Exon shuffling** by which recombination between domains from different genes originate a new gene form, an example of this case would be the *jingwei* gene in *Drosophila* (Long and Langley, 1993).

- **Gene duplication**, which is the most frequently cited mechanism of generation of new genes.

- **Retroposition**. This mechanism creates a duplicate of a gene at a different genomic location by reverse transcription.

- **Recruitment of sequence from transposable elements** (TE) by host genes as part of exonic sequences. In the human lineage the highest level of exonization of TEs corresponds to Alu elements (Sela *et al.*, 2007).

- **Lateral gene transfer**. In this case, a gene is transmitted from one organism to another from a different species has mainly been reported in prokaryotes so we will not discuss it further.

- **Gene fusion/fission**, in this case two genes become fused forming one single gene, or one gene becomes split in two genes, this has been reported mainly in prokaryotes, although a few cases in eukaryotes have also been identified (Cusack and Wolfe 2007, Frohlich *et al.* 2001).

- ***De novo* origination (exaptation)** of a gene from a previously non-coding genomic area (Levine *et al.* 2006, Begun *et al.* 2007).

In many real cases, as often occurs in biological studies, is not always clear in which of these categories a gene should be placed, as it may have originated by a combination of several of these processes. A case in point is the *jingwei* gene of *Drosophila.* This gene was formed when a mRNA from the alcohol dehydrogenase (*Adh*) gene was retroposed into the gene *yande*, between two of its coding exons, shortly after this gene had originated by a duplication of the *yellow-emperor* gene (Long and Langley 1993, Long *et al.* 2003). The resulting gene *jingwei* is a chimera formed by the first three exons of *yande* and the complete coding region of the *Adh* gene. Because of this complex origin this gene shows features of the exon shuffling process, by which different domains are created from the recombination of the previously existing *yande* and *Adh* domains, complete gene duplication, as this gave rise to the *yande* gene, and retroposition, as this is the process that inserted *Adh* into the *yande* gene.

Also the same type of process can sometime create different types of duplications. Gene duplication, exon shuffling and gene fusion/fission can all be caused by the duplication of a stretch of DNA which is subsequently copied and reinserted into the genome in a different location. Depending on the content of this DNA fragment and the location in which it is inserted the result can be any of these three processes.

Gene duplication

Gene duplication was suggested as the main source of new genes in the genome by Ohno (1970), and this has been verified in several studies (Maere *et al.* 2005, Katju and Lynch 2006, Scannell *et al.* 2007 among others). Genes may be duplicated by tandem duplication, segmental duplications that may affect several genes, or by larger duplications affecting complete chromosomes or genomes. In the case of smaller duplications, the main processes responsible for them are segmental duplication (also known as duplicative transposition) and reverse transcription of mRNAs (retropositions). Whole genome duplications and chromosome duplications are rare events, usually followed by a rapid loss of duplicate genes (Scannell *et al.*, 2007) and may be the cause of large lineage expansions, as has been suggested for the duplications at the origin of the vertebrates and of the teleost fish lineage (Cresko *et al.* 2003, Hoegg *et al.* 2007). These large duplication events are common in plants (Muller 1925, Prince and Pickett 2002). In animals, though less frequent, cases of polyploidy have also been observed in spiders (Schwager *et al.*, 2007), the salmonid fish (Johnson *et al.*, 1987) as well as other teleost fish (Amores *et al.* 1998, Jaillon *et al.* 2004), and in at least one species of mammal (Gallardo *et al.*, 2003).

Segmental duplications are duplications of large segments of genomic DNA that vary in size from 1 to more than 200 kb. These duplicated regions may occur in tandem or be relocated to other regions in the genome. The distribution of these duplications in the human genome is non-uniform between the different chromosomes and within each chromosome these duplications are more frequent in the peri-centromeric and sub-telomeric regions (Samonte and Eichler, 2002). It has been estimated that $\approx$ 5% of the human genome is formed by recent segmental duplications with a sequence identity $\geq$ 90% (She *et al.*, 2004). The mechanism of origin of these duplications is unclear, but the fact that most of them are interspersed throughout the genome argues against unequal crossing-over during meiosis as their primary mechanism of dispersal. Many of these duplications seem to cluster with other unrelated segmental duplications which seems to support the presence of certain areas in the genome more susceptible to incorporate these duplications when they occur (Samonte and Eichler, 2002).

The origin of many gene families has been associated to the process of gene duplication. Some well known examples of this are the hemoglobin and immunoglobulin families. Other cases such as the chorionic gonadotrophin (CG) have also been observed more recently and more cases are likely to be found thanks to the increasing availability of genomic information from multiple species.

The globin superfamily is a classic example of gene duplication followed by retention of the original function as well as neofunctionalization and nonfunctionalization. The hemoglobin molecule is formed by four peptides: two α-globins and two β-globins. There are several genes encoding different types of α and β-globin peptides. All of these genes as well as the myoglobin gene and several related pseudogenes have originated from the duplication of a single ancestral gene (figure 1.2 overleaf). Different combinations of these peptides can alter the properties of the resulting protein which results in different kinds of hemoglobin molecules having different affinities for oxygen depending on the combination of α and β subunits that form it. This ability to use different subunits in order to change the molecular properties of the hemoglobin has been exploited in some species of high flying birds that have several types of hemoglobins cir-

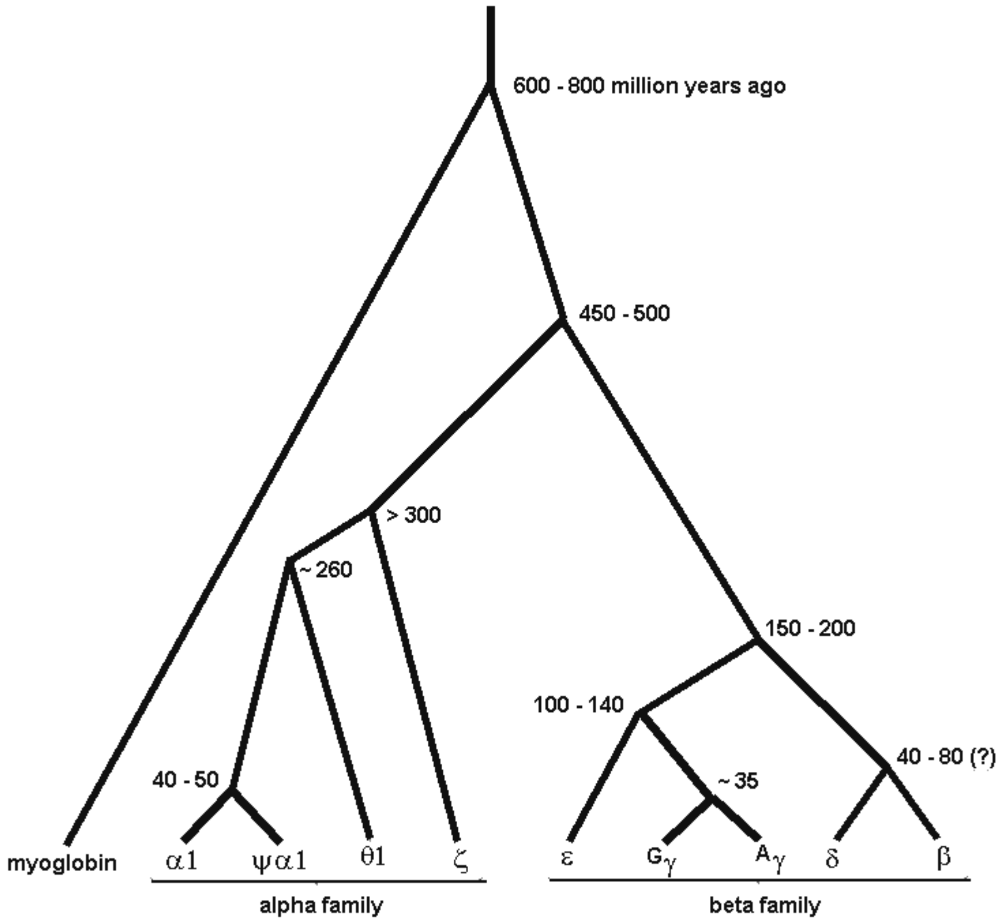


Figure 1.2: Relationship between the human globin genes and pseudogenes. Adapted from `http://nitro.biosci.arizona.edu/`

culating each of them with a different combination of subunit chains that have different oxygen affinities. This allows them to cope with the low oxygen levels encountered at high altitudes (Hiebl *et al.* 1988, Liu *et al.* 2001). Another way in which evolution has exploited these duplicates in human is by the use of a different combination of globin chains in the fetal hemoglobin. This fetal form has a higher affinity for oxygen allowing it to capture oxygen transported by the adult hemoglobin present in the mother.

Another example of a largely expanded group of genes formed by gene duplication is the immunoglobulin family of genes. This family includes more than 150 different genes that can be grouped into four different classes. The peptides generated by these four types of genes can be combined into a huge number of unique proteins.

A more recent example of gene duplication, both in origin and discovery, is the origin by duplication of one of the components of the chorionic gonadotrophin CG hormone in primates. This hormone is a fundamental signal in establishing pregnancy in humans and some other primates. It is formed by two subunits, α and β. The CGβ subunit, which confers the hormone its biological specificity is expressed only in placenta. This subunit originated by the duplication of the β subunit of the closely related luteinizing hormone (LH) in the Haplorrhini after the divergence of the tarsier lineage. These two genes, CGβ and LHβ differ mainly by the deletion of a single nucleotide eight amino acids upstream from the stop codon in the CGβ gene which changed the reading frame adding 24 aa from what would have been the UTR of the LHβ subunit. After this event this new gene has subsequently been duplicated repeatedly in the different groups of anthropoids (Maston and Ruvolo, 2002)

Duplicate genes seem to arise quite frequently in eukaryotic species (Lynch and Conery, 2000). In fact, up to 5% of the human genome is formed by segmental duplications that have originated in the past 35 My and are likely to have originated many new duplicates of existing genes (Eichler, 2001). However, not all genes that are duplicated will be maintained in the genome. After a gene duplication event, regardless of the process by which the original gene became duplicated the two copies may follow three different evolutionary paths (Force *et al.* 1999, Lynch and Conery 2000, Prince and Pickett 2002).

- Loss of one of the copies through accumulation of degenerative mutations.
- Preservation of both copies due to beneficial mutations in one of the copies while the other maintains the original function.
- Preservation of both of the copies due to the accumulation of deleterious mutations on both copies of the gene in a way in which neither of the duplicates is capable of maintaining the functions of the original gene by itself, but the presence of both copies will. This subfunctionalization model proposed by Force *et al.* (1999) is described in greater detail in chapter 5 on page 133.

A particularly larger number of duplication events have been observed in the lineage leading to mammals when compared to other lineages (Huerta-Cepas *et al.*, 2007) although this may be an artifact caused by the greater efforts invested in the study of mammals, which is reflected in the availability of much more sequence data from this lineage. Although gene duplication is frequent not all genes appear to have the same chance of becoming duplicated (Shakhnovich and Koonin, 2006). If gene families are partitioned into those containing at least one essential gene (E-families) and those that do not (N-families), the E-families are subject to stronger purifying selection. Because of this, genes in these families last longer in the genome and are more likely to undergo duplication with subsequent subfunctionalization or neofunctionalization than those with no essential gene. These evolve faster and are generally involved in fewer functions (less complex), so if duplicated they are more likely to end as pseudogenes. This means that belonging to an E or a N-family is a strong determinant of the future of a gene. The difference between the paralogues belonging to an E-family are greater than between those belonging to an N-family, probably because their longer average survival time gives them more time to differentiate.

One of the main advantages gene duplication and subsequent subfunctionalization offers to an organism is the possibility of separating the different function of a complex gene into separate genes. This will allow the duplicated genes to further specialize in the functions they retain without affecting the other functions which will be maintained by the other copy. An example of this can be seen in the subfunctionalization of the switch controlling the galactose use pathway in yeast. The ancestral gene form, which is present in *Kluyveromyces lactis*, performs both the galactokinase and co-induction functions. This ancestral gene was duplicated in the lineage leading to *Saccharomyces cerevisiae* and each of the two resulting genes (*GAL1* and *GAL3*) specialized in one of the original functions becoming more efficient in combination than the ancestral gene is on its own (Hittinger and Carroll, 2007).

Reverse transcription

Reverse transcription and integration in the genome is another mechanism by which a gene can be duplicated. In this case a gene is transcribed to RNA, spliced and a poly A tail is added, after which it is reverse transcribed to DNA and integrated into the genome. In most cases this will generate a dead-on-arrival pseudogene because it will lack the necessary regulatory sequences for its expression, or contains mutations in the sequence caused by the inaccuracy of the reverse transcription process that will render it inactive regardless of the location at which is integrated in the genome (Graur and Li, 2000). However in some rare cases a correct copy may be inserted close to another gene promoter, and in that way be expressed. This process would allow it to gain different expression profiles, and because of this in some cases different functions (Gregory, 2005).

An example of a gene that originated by retroposition is the human glutamate dehydrogenase GLUD2. This gene is a brain-specific isotype that originated from the retroposition of a copy of GLUD1, which is the original isotype and is involved in housekeeping functions. This gene originated in the common ancestor of the Hominoidea, originating after the split of this group from the Old World Monkeys (Burki and Kaessmann, 2004).

In a recent study Potrzebowski *et al.* (2008) show that many X-linked genes that have been retroposed to the autosomes since the common mammalian ancestor. This process has generated intronless autosomal copies of genes originally present in the X chromosome that are expressed in males during the period of meiotic sex chromosome inactivation. This migration of genes appears to compensate for the silencing of their X chromosome gene copies during and after the meiotic phase of spermatogenesis (Potrzebowski *et al.*, 2008).

Exon shuffling

Exon shuffling or domain shuffling, as it is also known, occurs when two or more exons from different genes are brought together or the same exon is duplicated to create a new intron-exon structure, duplicating an existing domain. The most frequent processes that lead to this are illegitimate recombination, which is the recombination of non-homologous genomic sequence, or retroposition.

Some well known examples of genes that have been formed by the combination of exons from different original genes are the previously mentioned *jingwei* (Long and Langley 1993, Zhang *et al.* 2004) and the fucosyltransferase gene FUT8. The fucosyltransferase genes belong to a family of transmembrane enzymes. They present a common structure that includes a short NH_2-terminal cytoplasmic tail, a signal membrane anchor domain, a stem region and a globular COOH-terminal catalytic domain. This family is formed in human by nine members, all of them with mono-exonic coding regions except for the $\alpha1,6$ fucosyltransferase (FUT8) gene. This gene includes additional exons that encode peptides which can be identified in exons from phylogenetically unrelated proteins (Javaud *et al.*, 2003).

Exon duplication can also have a big impact in the evolution of eukaryotic genes. In animals up to 10% of the genes in human fly and worm show tandem exon duplications (Letunic *et al.*, 2002). Examples of this process can also be found in plants such as the case of the TCH3 gene in *Arabidopsis thaliana* (Sistrunk *et al.*, 1994). This gene is a calcium binding protein induced by touch and darkness and we will examine it in more detail in chapter 6 on page 171.

Gene fusion/fission

Two adjacent genes can be fused into a single transcript by read-through transcription caused by the deletion or mutation of the transcription termination signal. Many cases of fusion have been identified in prokaryotes where in most cases the fused genes will be also functionally linked before the fusion event (Yanai *et al.*, 2002). These events are more frequent than fission events, which is not surprising, as unless the two resulting proteins formed by a gene fission event are able to function independently they will not be maintained (Kummerfeld and Teichmann, 2005). Another source of gene fusion that has been recently suggested in eukaryotes is tandem chimerism. This occurs when alternative splice variants are formed that contain exons from two adjacent genes, but each of this genes is completely functional with complete promoter and transcription termination regions (Akiva *et al.*, 2006). Evidence of tandem chimerism has been reported for a significant fraction of tandem pairs in the human genome although the functionality of many of these transcripts is still uncertain (Parra *et al.*, 2006). An example of a gene fusion can be found in the SPAG11 gene (originally described as EP2). This gene is involved in the immune response as well as several signal pathways and may be important for sperm maturation (Yenugu *et al.* 2006, Hall *et al.* 2007). This gene was formed by the fusion of two ancestral β-defensins and could be classified as a case of tandem chimerism, as it shows alternative transcripts that include both of the ancestral genes and also transcripts that span only one of them (Frohlich *et al.*, 2001). We will examine this gene in more detail in chapter 5 on page 133.

Transposable elements

Exonization of DNA from TEs, particularly Alu elements has also been described, and is a source of new exons and alternative splice variants in many eukaryotic genes. These new exons are frequently alternative exons with a low rate of inclusion in the transcript, and this may allow for their high frequency if their presence does not significantly disrupt the function of the major transcript (Sela *et al.*, 2007). The double-stranded RNA-specific editase 1 (RED1/ADAR2) is an example of this phenomenon. This gene shows two alternative transcripts, the longest of them includes an Alu-J cassette in the deaminase domain. Both of these alternative proteins show the same substrate specificity, however, they differ in their catalytic activity, which is higher in the shorter form (Gerber *et al.*, 1997).

Exaptation of non-coding regions

De novo generation of genes from non-coding sequence has been described only in a few cases (Brosius, 1999), however, there are many different ORFs distributed along the genome of every organism that do not form part of coding regions, and if any of these ORFs acquired the necessary signals for expression they could theoretically originate a new protein. The process of exaptation of non-coding regions may be more frequent in the case of partial gene formation from non-coding regions or regions that are part of genes encoded in the opposite strand originating overlapping genes (Makalowska *et al.*, 2007). Binding sites of transcription factors and promoter sequences are generally only a few nucleotides in size (Balding DJ., 2003), so it would not be inconceivable that they could be formed by random mutation and promote the transcription of nearby regions within the genome. Some TEs such as LTRs contain both promoter and enhancer elements and their transposition to an area upstream from an existing ORF could in theory activate its expression (Brosius, 1999). There is a large amount of non-coding DNA which is expressed in mammals (Mockler *et al.*, 2005), up to 70%

of the genome is transcribed, and in most cases there is evidence of transcription from both strands. A large fraction of this transcription occurs in a differential way between tissues or developmental stages, and could be potentially functional (Pheasant and Mattick, 2007). As a large fraction of the genome is transcribed, one of the many ORFs that exist in the genome may be incorporated to one of these transcripts and translated.

Regardless of the manner in which a non-coding sequence is exapted into a new gene, it is difficult to detect these cases. This is because if a gene originates *de novo* the sequence from which it originated is likely to have mutated beyond recognition in the more distant lineages, and even if the species to which it is compared is sufficiently close for the sequence to still be recognizable it will likely be labeled as a pseudogene.

Because the three primate species are sufficiently close for the non-coding areas to be highly conserved, they are an ideal group to use for a comprehensive search for *de novo* gene formation by exaptation of a non-coding region. This is examined in detail in chapter 4.

Gene loss

The gene content of an organism can also decrease. There are fewer mechanisms that will cause reduction in the number of genes present in the genome. The main causes of decrease in the gene content are deletions that will remove a sufficient fragment of the gene to cause its inactivation and deleterious mutations that also inactivate the gene. Large deletions that involve significant fractions of chromosomes may also occur, although in most cases these will not produce viable individuals, as they are more likely to include essential genes.

When a gene is inactivated if it is not essential for the survival or reproduction of the organism, the sequence will rapidly degenerate due to random mutation of its sequence as it is no longer under purifying selection. If the gene was necessary for the survival or reproduction of the individual, the individual will produce no offspring and the mutant form will disappear from the population.

In some cases the elimination of one gene may lead to the inactivation of other functionally related genes, presumably because they may become redundant if a fundamental gene for the process in which they are involved is removed. This co-elimination of functional pathways has been observed in comparisons between *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (Aravind *et al.*, 2000).

Gene loss has also shaped the caspase family in mammals. This group of genes encode a group of proteases that play a central role in processes such as apoptosis, inflammatory pathway signaling and immune cell proliferation among others. Caspase 12 is inactivated in the majority of the human population (Wang *et al.*, 2006). This loss appears to be a case of the less is more hypothesis suggested by Olson (1999), as the loss of this caspase shows evidence of positive selection in the human population. Caspase 17 was also lost in the lineage leading to therian mammals (marsupials and placental mammals) and caspase 18 was deleted in the lineage leading to eutherian mammals (Eckhart *et al.*, 2008).

Gene loss is more difficult to identify than gene gain by duplication, as in order to identify a loss the compared genomes must be complete or reasonably covered. However the large increase in available whole genome sequences will likely result in many more studies focusing on gene loss patterns between different species.

1.3.3 Genome structure

The information contained in the genome in all organisms is organized in a linear fashion from the nucleotide level to the chromosomal level. This information is stored at the nucleotide level within the order of the four different nucleotides along the molecule, and at higher levels within the order of different motifs that will produce certain structural patterns that allow the recognition of the different regions of the helix by different molecules. The double helix structure of the DNA molecule implies, as was noted by Watson and Crick, the manner in which the genetic material is replicated (Watson and Crick, 1953). By which the double strand of DNA would be separated and each of the single strands would act as templates for the synthesis of a new complementary molecule. The result is two identical copies of the original DNA molecule. This manner of replication means that not only the complete gene content is copied, but also the order in which all the different elements appear in the genome.

In prokaryotes the genetic information is organized into a circular DNA molecule, whilst in eukaryotes the genetic information is split between a set of chromosomes each of them a single linear molecule of DNA. Although the organization into linear chromosomes poses some problems that do not appear in the case of circular chromosomes, the essential manner in which the replication occurs is the same. Each chromosome will give rise to two identical copies of itself each of them containing one strand of the original double helix and one newly synthesized strand.

The genome of any two organisms originates from the genome of their common ancestor, so the same as we expect a conservation of the nucleotide sequence within a gene that appears in both of these species we would expect a conservation of the gene sequence within the genome. Although during meiosis in diploid organisms there is a recombination between the homologous chromosomes, these events do not alter the order of the genes along the chromosome.

However, there are different rearrangement mechanisms that can alter the gene order within the chromosome or move fragments from one chromosome to another without involving sequence duplication. Those mechanisms that alter the gene order within a chromosome are called intrachromosomal rearrangements and can be transposition events by which a fragment of the chromosome is excised and moved to another location within the same chromosome, in the same or inverted orientation, or in-place inversions, by which a fragment of the chromosome is excised and reinserted in the same position but with an inverted orientation.

The mechanisms that rearrange the gene content within or between chromosomes are intrachromosomal and interchromosomal rearrangements respec-

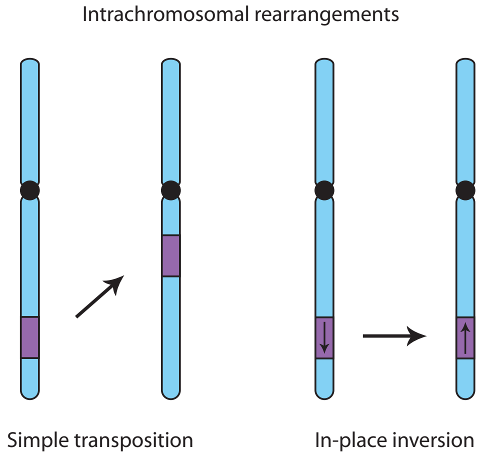


Figure 1.3: This figure shows a schematic representation of the two major types of intrachromosomal rearrangements. Modified from (Gregory, 2005)

tively. Interchromosomal rearrangements can be reciprocal translocations, these are the most common by which two fragments from two different chromosomes are exchanged, simple translocations, by which a terminal fragment from one chromosome is translocated to another chromosome, with no reciprocal sequence being translocated back, and intercalary translocation in which a fragment of one chromosome is translocated to a non-terminal part of another. A specific type of reciprocal translocation which occurs when the long arms of two acrocentric chromosomes, where the centromere is close to the end of the chromosome, fuse together is called Robertsonian translocation.

As well as these mechanisms which do not involve significant gain or loss of sequence, some of the processes described previously such as segmental duplications can produce rearrangements that also involve significant gain of sequence. Large deletions can also occur which will remove any genes contained in the affected region. Because of these mechanisms the gene order conservation between two related species will decrease with their phylogenetic distance. The occurrence of segmental duplications could lead to a subsequent chromosomal rearrangement, as it will allow the homologous recombination between the two

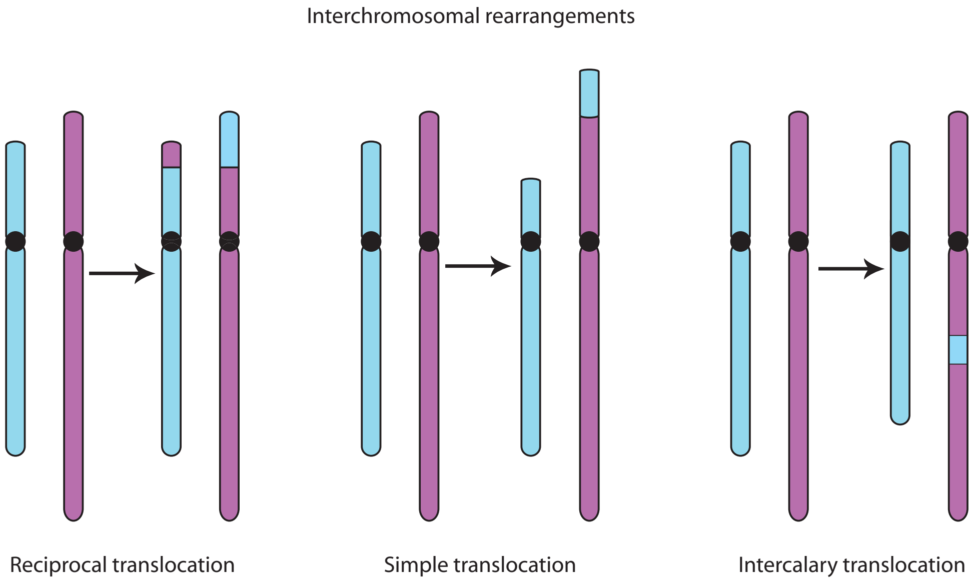


Figure 1.4: This figure shows a schematic representation of the three major types of interchromosomal rearrangements. Modified from (Gregory, 2005)

duplicated areas which may be located in different chromosomes. Highly homologous sequences are more likely to undergo unequal recombination which may lead to large-scale chromosome rearrangements, including deletions, duplications pericentric inversions and translocations (Samonte and Eichler, 2002). In fact, segmental duplications were identified in the breakpoint regions of six of the nine pericentric inversions that distinguish human and chimpanzee genomes (Kehrer-Sawatzki and Cooper, 2007b).

Chromosome fusions and fissions, although less frequent, may also occur. A well known example of this can be found in human chromosome 2, which is formed by the recent fusion of two ancestral primate chromosomes.

Chromosomal rearrangements can have an important effect on speciation. They can create genetic barriers and prevent recombination. This effect of chromosome rearrangements on speciation was elegantly demonstrated by Delneri *et al.* (2003) when they produced viable spores from *Saccharomyces* "*sensu stricto*" hybrids. These *Saccharomyces* species are capable of mating but produce sterile hybrids. Restoring the co-linearity of their chromosomes produced fertile hybrids.

1.3.4 Gene order

Genome organization traditionally has been studied using cytogenetic mapping and genetic-linkage, and more recently chromosome painting and radiation hybrid techniques allowed chromosomal homologies to be visualized between different species, however these techniques will not allow the identification of small rearrangements or changes in the gene order within a conserved block.

Synteny –from *syn*, together and *tainia*, band or ribbon – is used to refer to genes that occur in the same chromosome. The degree of synteny conservation between two organisms will reflect the number of interchromosomal rearrangements that have occurred between them since the two species diverged.

Synteny maps have been created between human and many other organisms, and with the increasing availability of high quality genomic sequences, information on gene order is proving very useful to determine small scale rearrangements and gene orthology in closely related species.

With the current genomic information it is increasingly possible to not only use the synteny information, but also the gene order information in order to determine the degree of conservation of the genomic structure between two organisms. This allows us to identify changes caused by intrachromosomal events as well as those produced by interchromosomal rearrangements at a far greater resolution than previously. This allows the identification of small scale rearrangements and segmental duplications that were invisible to previous techniques and provides evidence of the frequency of these small scale rearrangements even when comparing the genomic sequence of closely related species. This demonstrates that small translocations, inversions and duplications are not uncommon in eukaryotic genomes (McLysaght *et al.* 2000, Vision *et al.* 2000, Murphy *et al.* 2005). An example of this can be found in chromosome 17, which is the largest human autosome with orthology to a single mouse chromosome, the distal half of mouse chromosome 11. This chromosome is rich in protein coding genes, having the second highest gene density in the genome, and is also enriched in segmental duplications. It also shows evidence of many intrachromosomal rearrangements compared to mouse and has accumulated many more segmental duplications (Zody *et al.*, 2006). Those areas of the genome located within 10 Mb of the end of the chromosome show a higher rate of divergence than those in other chromosomal locations.

Results from previous studies suggest breakpoints within the mammalian genomes are not randomly distributed. Evolutionary recombination rates within genes are low (McVean *et al.*, 2004), however in gene rich regions they are the highest, this may indicate a high level of of recombination around loci that are under selective pressure is beneficial as it allows for selection to act independently on the different genes (Murphy *et al.*, 2005). Nearly 20% of chromosome breakpoint regions were reused during mammalian evolution (Pevzner and Tesler, 2003). These reuse sites are enriched for centromeres and telomeres. Segmental duplications appear on most primate specific breakpoints, and often flank inverted chromosome segments. Because not all segmental duplications are accompanied by chromosomal rearrangements, but most of the rearrangements (98%) contain segmental duplications it is likely that these duplications promote non-allelic homologous recombination and thus the chromosomal rearrangements (Murphy *et al.*, 2005).

When compared to *Tetraodon nigroviridis* the human genome shows a larger number of rearrangements from the reconstructed ancestral genome (Jaillon *et al.*, 2004). Within the mammalian group rodent genomes have a very high rate of rearrangement when compared with other mammals (Murphy *et al.*, 2001), and many of the observed cases of breakpoint reuse occurred within this lineage (Murphy *et al.*, 2005).

In the case of primates the frequency of rearrangements seems to follow the general mammalian trend, being quite conserved, with only one major exchange every 10 My. Reconstruction of the ancestral primate genome shows that at the resolution of chromosome painting 19 out of the 26 ancestral primate chromosomes have been conserved in human since the divergence of the primate lineage 60-70 Mya (reviewed by Murphy *et al.* 2001).

1.4 Intron Evolution

Introns are stretches of intragenic sequence that are transcribed into RNA, but are removed from the mRNA before it is translated to protein. There are four major classes of introns: self-splicing groups I and II introns, tRNA and/or archaeal introns and spliceosomal introns (Rodriguez-Trelles *et al.*, 2006).

1.4.1 Self Splicing introns

Self splicing introns fold into a ribozymic structure which catalyzes their own excision, and may propagate into new sites through reverse transcription assisted by proteins they encode themselves. These introns are divided into two classes Group I and Group II (figure 1.6 on page 29). Group I appear in bacterial and organellar genomes, and in the ribosomal RNAs (rRNAs) of some protists and fungi. Group II introns appear in bacteria, and organellar genomes of fungi, plants and protists. They have not been observed in nuclear genomes or animal mitochondria. These introns share some similarities with spliceosomal introns that may suggest an evolutionary relationship. However there are also arguments that can be raised against this as some of the eubacterial group II ”introns” lie between and not within genes, so they are not real introns, and the structural similarities may be caused by constraints on how splicing can be carried out, and not by a common origin (Lynch and Richardson, 2002).

Transfer RNA (tRNA) introns are found in the nuclear tRNA genes of eukaryotes and also in messenger RNA (mRNA), rRNA and tRNA in archaea. This type of introns are removed through a completely different pathway from that of

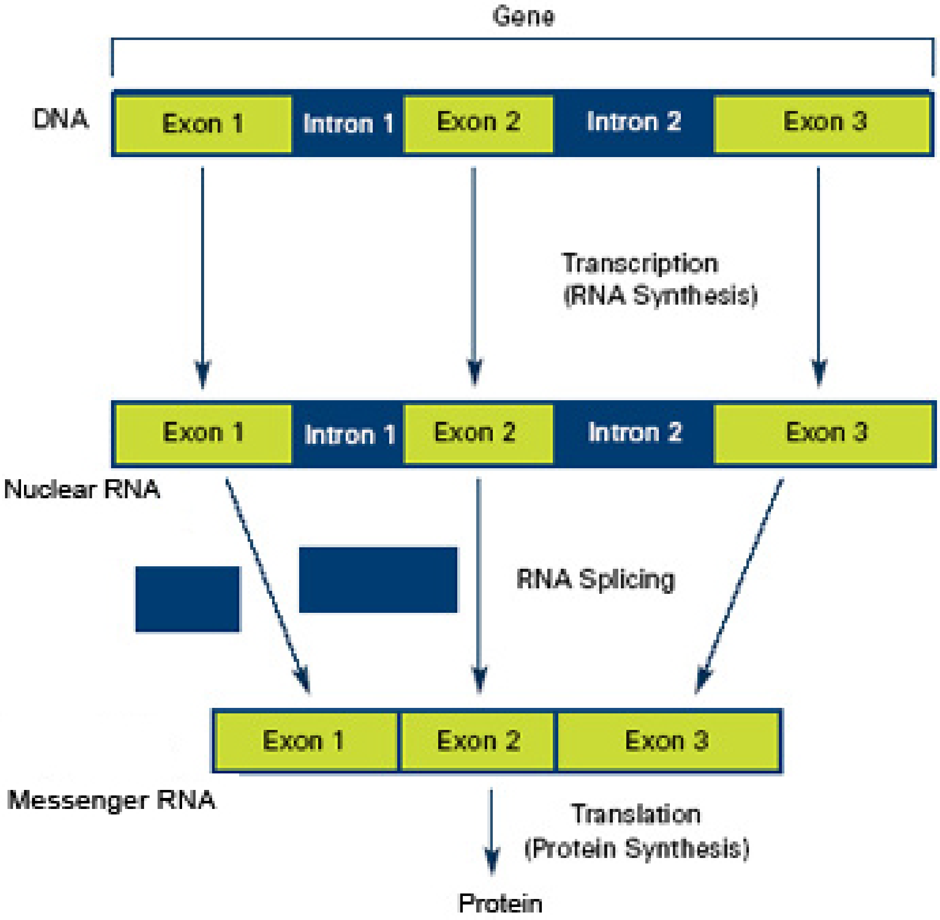


Figure 1.5: Introns present in eukaryotic genes. Modified from `http://publications.nigms.nih.gov/thenewgenetics/chapter1.html`

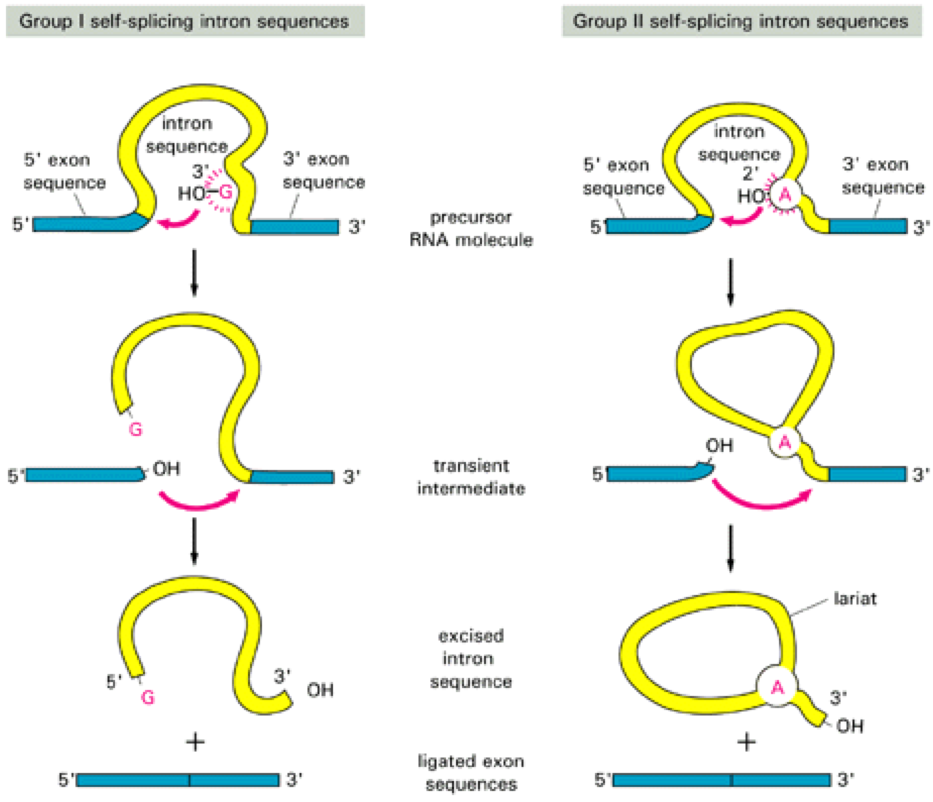


Figure 1.6: Schematic representation of the splicing mechanism in Group I and Group II introns. Group II introns form a lariat similar to that formed during the splicing of spliceosomal introns. Modified from (Alberts *et al.*, 2008)

spliceosomal introns.

1.4.2 Spliceosomal introns

Spliceosomal introns appear in the nuclear genome of all eukaryotes, and are the least conserved at the sequence level. Splicing is the process by which introns are removed from the pre-mRNA. The splicing process begins with the recognition of specific sequences or splice sites located at the intron-exon boundaries and the assembly of the splicing machinery or spliceosome, which is composed of 5 small nuclear RNAs and more than 50 proteins, at these sites.

There are two kinds of spliceosome, the major or U2 spliceosome, and the minor or U12 spliceosome. The U2 spliceosome deals with the classical introns which start with GU and end with AG. The U12 deals with most of the non-canonical introns, and has only been found in some eukaryotic lineages. Although most of the spliceosomal proteins are shared between them, four of the five core snRNAs are not (Lynch and Richardson, 2002).

As well as the splice sites other RNA sequence elements, the exonic and intronic splicing enhancers (ESEs and ISEs) as well as the exonic and intronic splicing silencers (ESSs and ISSs) play a role in the recognition of splice sites. These splice site recognition sequences can be grouped into two different types depending on if the recognition signal is located mainly within the intron (intron definition) or mainly within the exon (exon definition). The exon definition is frequent in organisms with very large introns, while the intron definition is more frequently used by those introns that have a small size (Jaillon *et al.*, 2008). Splicing occurs mainly within the nucleus, which allows spatial and temporal separation between the transcription and the protein synthesis which is not available in the case of prokaryotes, where protein translation may begin before the transcription of the gene has finished. Splicing outside of the nucleus has been described in platelet activation (Blaustein *et al.*, 2007), but is likely to take place only in certain specialized cells.

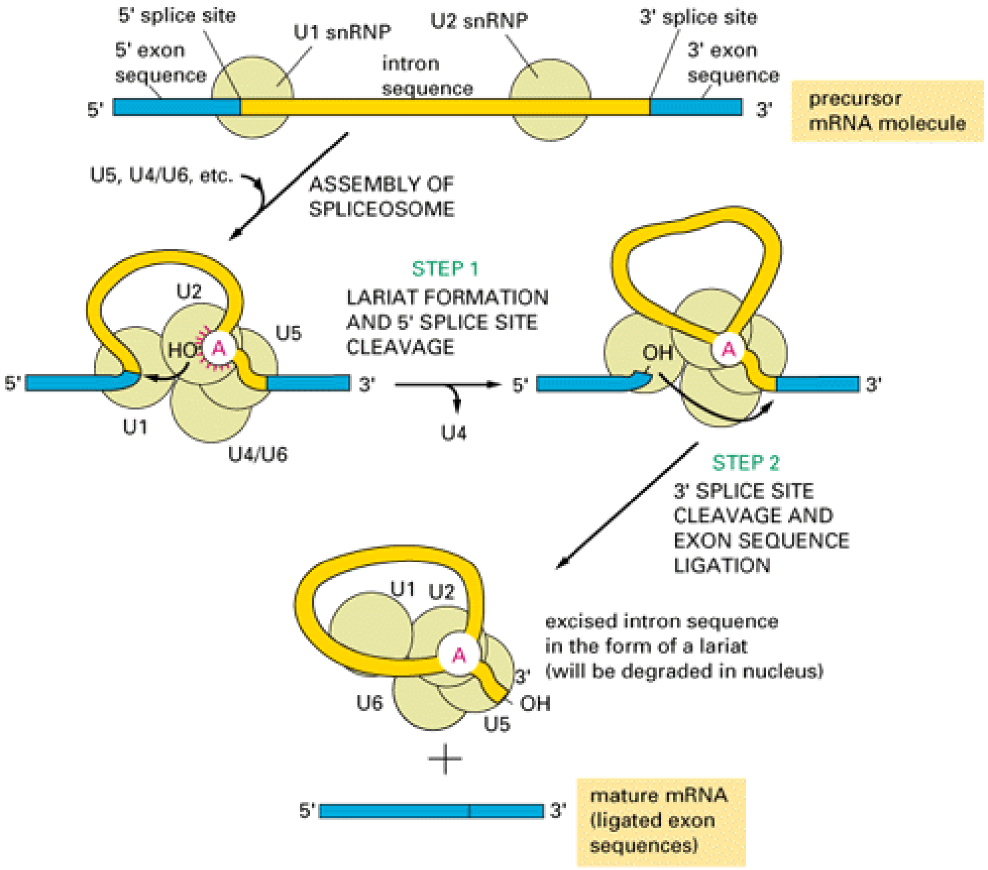


Figure 1.7: Schematic representation of the splicing mechanism in spliceosomal introns. Modified from (Alberts *et al.*, 2008)

1.4.3 Biological significance of introns

Because of the weak splicing signals present in many of the small introns (at least), there is an additional proofreading mechanism in eukaryotes, the nonsense mediated decay (NMD) pathway by which those incorrectly spliced mRNAs are degraded.

Different functions have been suggested for introns which may explain their persistence during evolution.

Their presence between exons has been suggested to allow increased recombination between coding exons (Rodriguez-Trelles *et al.*, 2006). The presence of introns also allows for the alternative usage of different exons from the same gene, which increases the number of proteins that can be produced without altering the number of genes present in the genome. This alternative splicing can be regulated in a tissue-specific or developmental stage-specific manner, and is possible due to the fact that in many eukaryotic introns the splicing signals are very degenerate and insufficient to achieve a precise splicing by the spliceosome. Correct identification of the splice sites is achieved by the low specificity binding of multiple splicing regulatory proteins that bind to the mRNA and aid the recruitment of the spliceosomal complex to the correct area. The activity of each of these proteins, as in the case of many other proteins within the cell depends on their phosphorilation state and cellular location. By varying these different proteins can be excluded from the available pool of active regulatory proteins that will aid the splicing reaction. In this way alternative selection of different splice sites can be achieved with a high accuracy and speed, as it does not require new synthesis of different regulatory proteins (Blaustein *et al.* 2007, Stamm 2008).

Several routes that regulate the alternative splicing of specific genes have been described in varying degrees of detail. These include growth factors, cytokines, hormones and depolarization, changes in the alternative splicing patterns can be cause d by different physiological stimuli, from glucose levels to stress (Stamm, 2002). Some of these alternatively spliced proteins are themselves involved in

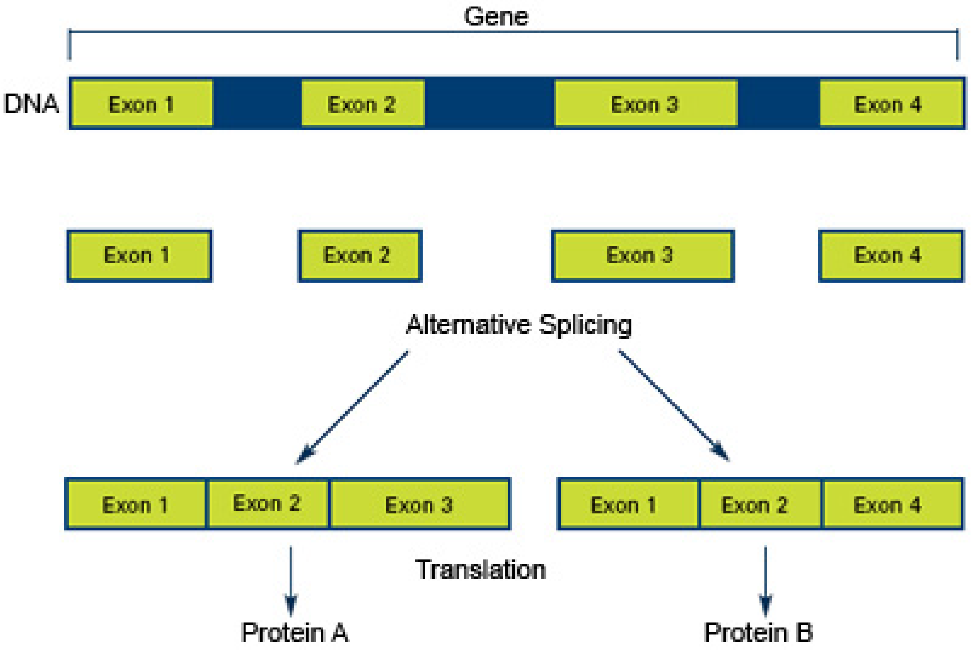


Figure 1.8: Schematic representation of the alternative splicing mechanism. Modified from `http://publications.nigms.nih.gov/thenewgenetics/chapter1.html`

intracellular signaling pathways, which makes alternative splicing an additional signal regulator in eukaryotes (Blaustein *et al.*, 2007).

Splicing patterns vary from cell to cell and can be rapidly changed in response to external stimuli. This is achieved by differential phosphorylation of the different proteins involved in the splicing mechanism as well as by changes in their sub-cellular distribution and *de novo* protein synthesis (Stamm, 2002).

Alternative splicing is together with gene duplication one of the major contributors to the proteome variability in the eukaryotic lineage (Blaustein *et al.*, 2007). The percentage of genes undergoing alternative splicing is higher when comparing vertebrates and invertebrate (Kim *et al.*, 2007). In the mammalian lineage it has been estimated that between 40% and 75% of the genes in the human genome possess at least two alternative splice variants (Babushok *et al.*, 2007), this allows for the number of proteins that are encoded in it to be much larger than the number of genes. This mechanism has been suggested as one of the explanations of the huge variation in complexity between eukaryotes that possess similar gene numbers (Xing and Lee, 2006), and may be responsible for a large fraction of the mammalian phenotypic complexity.

1.4.4 Origin of spliceosomal introns

The origin of introns is a controversial issue since their discovery. The two main traditional hypotheses that have been suggested are the "introns early" and the "introns late", and additional "introns first" theory has also been suggested more recently.

According to the “introns early” hypothesis introns are an ancestral feature, and the absence of introns is a derived feature. This theory suggests that the last universal common ancestor (LUCA) possessed a highly redundant genome which allowed it to prevent information loss in a primitive high error rate replication system. The genes would be coded by small stretches of DNA separated by introns that would allow recombination between the different pieces in order to maximize the possibility of assembling functional proteins. These introns became increasingly redundant as the replications mechanisms increased in efficiency, and were eventually lost in prokaryotes. However, in eukaryotes, by chance they acquired new functions, and were retained. Splicing has also been hypothesized to be a molecular relic of this earliest mode of recombination which may have taken place in the RNA world to allow the excision and ligation of independent fragments of RNA into functional fragments in face of a rapid information decay (Rodriguez-Trelles *et al.*, 2006). Another hypothesis that is radically opposite is the "introns late" hypothesis. According to this, the ancestral eukaryote had no introns, and those organisms that possess introns have acquired them since. If introns were mobile elements that spread via replicative transposition they could account in part for the abundance of these sequences in eukaryote genomes. This theory was supported by the discovery of self-splicing group II introns, which share similar mechanisms with spliceosomal introns and are capable of transposing into different parts of the genome (Dickson *et al.* 2001, Rodriguez-Trelles *et al.* 2006).

A more recent hypothesis is the "introns first" theory. This theory accommodates for some new observations that have been possible with the large amount of available genomic dada. This increased availability of genomic data has allowed comparative analysis of spliceosomal introns across the major eukaryotic groups. The results indicate intron turnover has been fast enough to remove most of the original intron-exon boundary distribution, and the mechanisms by which introns are gained and lost vary between the different lineages. Some other results point to a very early origin of the spliceosome, before the most recent common ancestor of all living eukaryotes. The fact that the spliceosome is possibly the largest RNA-protein complex in the eukaryotic cell, being larger than the ribosome, is consistent with this early origin model in which it evolved gradually over a long period of time, in an intron rich genome. The introns first theory assumes modern organisms descend from a primitive RNA world in which enzymatic reactions were catalyzed by ribozymes, and the first proteins were RNA binding chaperone like proteins. Small nucleolar RNAs (snoRNAs) are required for the maturation of rRNAs, these are required for protein synthesis, and thus snoRNAs should predate proteins. As most snoRNAs occur in introns of chaperone-like proteins, at least these introns must be older than the exons that surround them. And if introns existed, a splicing mechanism must have been available in order to splice them. In this scenario exons would have originated from the unused portions of RNA between the introns which would have originally been type-I or type II self splicing introns (Rodriguez-Trelles *et al.*, 2006). Intron encoded miRNA could originally have appeared to regulate these intronic genes and later the complete mRNA that was produced (Lin *et al.*, 2006). The original intronic genes would have been self recombining introns, that catalyzed their own excision and recombination, helping to increase the functional size of the self-replicating molecules.

Although there currently seems to be little doubt the eukaryotic ancestor contained introns, it is less likely that prokaryotes contained introns (Lynch and Richardson, 2002). Regardless of the precise origin of introns, most steps of mRNA processing in modern eukaryotes depend on the splicing machinery, to a point in which few modern eukaryotes if any would be able to survive without introns as many processes within the protein synthesis rely heavily on the assembly of the spliceosome if not on the splicing directly (Lynch and Richardson, 2002), one example of this is the NMD pathway mentioned earlier which allows the monitoring of prematurely ending transcripts preventing the synthesis of potentially deleterious proteins.

A more recent hypothesis was suggested by Martin and Koonin (2006). This hypothesis also suggest an early origin of introns in eukaryotes, but for a different reason. It proposes a causal relation for the observation that all known organisms that posses introns posses mitochondria (at least originally) and also posses a nucleus. This theory suggests that the mitochondrial ancestor, an α-proteobacteria symbiont was acquired as an endosymbiont by a prokaryotic host that also lacked a nucleus. The endosymbiont probably contained group II introns. These group II introns may have migrated from the mitochondrial genome to the genome of the host, and subsequently spread within it to different positions as transposable elements. Degeneration of some of these group II introns that prevented their correct excision could occur, which was likely to have deleterious effect. Splicing may still be possible in *trans* for these introns with the help of elements provided by other intact group II introns. However, protein synthesis is faster than splicing, and ribosomes, competing with these splicing elements in the cytosol, would synthesize proteins from un-spliced mRNAs. In this scenario any organism that developed a spatial separation between the splicing machinery and the translation process that prevented this from happening would possess a significant selective advantage by only translating those mRNAs that had been correctly spliced (Martin and Koonin, 2006).

1.4.5 Intron gain and loss

There is much more evidence of intron loss than gain, and many different mechanisms have been proposed for each of these phenomena, but few unequivocal examples have been found. The simplest mechanism for intron loss is deletion, which does not require precise removal, as long as the remaining nucleotides maintain the reading frame and do not introduce a stop. Reverse splicing has been suggested as a possible mechanism of intron gain, as this would ensure the new introns can be spliced (Lynch and Richardson, 2002). Identifying the mechanisms by which recent intron gain and loss has occurred within different genes is an important issue for understanding the recent evolution of gene structure. Some of these processes that can result in the gain or loss of introns are the same that can also produce the gain and loss of complete genes. This is the case for exon shuffling and retroposition of intronless mRNAs. A detailed summary of the known mechanisms of intron gain and loss can be seen in chapter 6 on page 171.

1.5 Comparative genomic studies between the sequenced primates

Hominoid evolution and the speciation of the human lineage are one of the most interesting topics in evolutionary biology. This stems in part from our position as members of this species, but another part stems from the desire to identify those changes that have allowed humans to occupy the unique position they hold today. This uniqueness is reflected in the fact that ours is the only species that has developed a written language, and thanks to this been able to increase the knowledge of the species as a whole beyond the learning capacity of a single individual.

Differences between the human genome and our closest relatives, as mentioned earlier, are assumed to be influenced strongly by lineage specific genomic changes. Search for some of these differences has given rise to many comparative studies (reviewed in Kehrer-Sawatzki and Cooper 2007a), including the present one.

The availability of the chimpanzee genome (CSAC, 2005) and more recently of that of the macaque (Gibbs *et al.*, 2007) has made it possible to compare these three species at a much higher resolution.

At the nucleotide level the difference between human and chimpanzee is quite small, between 1 and 2% in the alignable regions, while differences with macaque are slightly larger, up to 7% difference in aligned regions.

At a higher level the microscopic structure of these three genomes is highly conserved (figure 1.9 overleaf). There are ten chromosomal rearrangements between human and chimpanzee, one of these is caused by the fusion of the ancestral chromosomes 2A and 2B to form the human chromosome 2. The other nine correspond to pericentric inversions. Two of them were identified as human specific in chromosomes 1 and 18, and the other seven as specific to the chimpanzee lineage by Gibbs *et al.* (2007). There are also 43 microscopic rearrangements between the reconstructed human-chimpanzee ancestral genome and the macaque

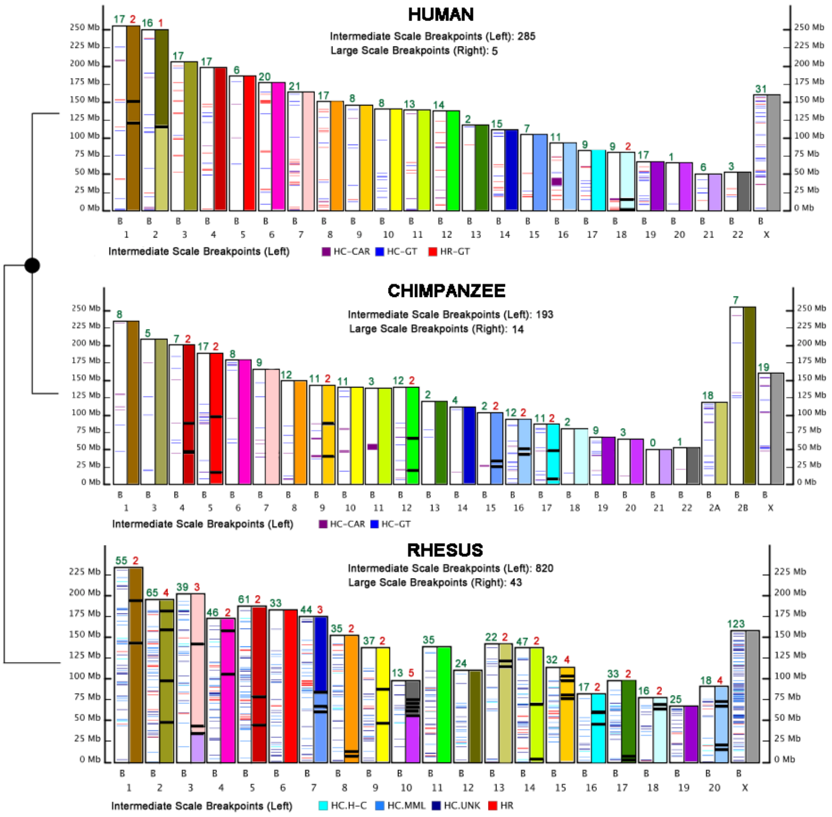


Figure 1.9: This figure shows existing breakpoints between macaque, chimpanzee and human. Chromosomes are represented by two bars. The white bar on the left shows thin horizontal lines that represent submicroscopic breakpoints. The bar on the right is coloured according to the homologous human-chimpanzee chromosome and shows thick black lines representing breakpoints that are visible at a microscopic scale. Modified from (Gibbs *et al.*, 2007)

lineage. In addition to these there are more than 1000 submicroscopic rearrangements that have occurred through these three lineages (Gibbs *et al.*, 2007). In addition to these differences many segmental duplications specific to human or to chimpanzee have been identified (Cheng *et al.*, 2005) as well as differences in the type and number of simple repeats and transposable elements present in the three species (CSAC 2005, Gibbs *et al.* 2007).

At the gene content level unambiguous orthologues for a large fraction of the human genes were identified from the initial chimpanzee sequence, and with the release of a higher coverage version in 2006 this fraction increased even more. A high number of 1:1:1 orthologues was also identified when comparing to the initial macaque genome (Gibbs *et al.*, 2007). However there are still many genes with no clear orthologue. Previous studies have reported humans specific gene loss events (Wang *et al.* 2006, Hahn *et al.* 2007). There are surely many cases of chimpanzee specific gene losses, indeed this species may posses even more pseudogenes than human (Wang *et al.*, 2006). However, the lack of a more refined sequence makes it more difficult to identify chimpanzee specific losses as their absence may be caused by their location in un-sequenced regions of their genome.

Gene expression changes are also an important source of phenotypic differences, and these have also been detected between human and chimpanzee. This demonstrates the speed with which large changes in gene expression may occur over short periods of time. These differences between species are greater in tissue specific genes (Kehrer-Sawatzki and Cooper, 2007a).

Searches for regions undergoing positive selection in human as indicators of those gene that may have conferred our species its fitness advantages have been carried out on many sets of genes (reviewed by Nielsen *et al.* 2007) and many genes that appear to be under positive selection in human, such as the forkhead box P2 (FOXP2) gene related to speech or variants of the lactase (LCT) gene, which allows the digestion of milk, that persist into adulthood, have been identified among others (reviewed by Sabeti *et al.* 2006), although studies of genes specifically expressed in the brain did not find any large adaptive changes (Shi *et al.*, 2006). Some of the difficulties in these studies will be to distinguish between those positively selected genes that are related to human specific evolution and those that are not such as the immune/defence related genes, as well as agreeing in which methods are the best to use, as many of these studies have identified different sets of genes as subject to recent selection and these sets show little overlap (Nielsen *et al.*, 2007).

1.6 Current genome sequencing and assembly techniques

Modern genome sequencing is based on the sequencing technique developed by Sanger *et al.* (1977). Since this original development the technique has been improved considerably and the speed, accuracy and length of the sequence reads that can be achieved are much better. Modern sequencing stations are capable of completing the sequence of an average sized bacterium in a day. However, the major problem when sequencing large regions remains the length of each sequencing read. Each of these reads will have a length from several hundred nucleotides up to, in the best of cases, a little over a thousand. Because of this in order to obtain the complete genomic sequence of an organism the genome must be fragmented and the results of the sequencing of each of these fragments reassembled in order to reconstruct the genome sequence. Assembly of the reads into longer fragments relies on identification of overlapping regions between fragments that can be used to link them together into longer stretches of continuous sequence. This is done using different genome assembly programs such as GigAssembler (Kent and Haussler, 2001) or the more recent Arachne (Batzoglou *et al.*, 2002) or PCAP (Huang *et al.*, 2003) among others. The number of times each individual nucleotide will be sequenced on average is the genome coverage, this means on a 10x coverage each nucleotide will have been sequence on average ten times. A higher coverage means more reads are likely to show significant overlap with others that will allow them to be linked together. This will result in a better quality of the resulting assembly with fewer gaps. However greater coverage also increases the cost.

The two main approaches used for whole genome sequencing are a hierarchical or clone based approach and a whole genome shotgun (WGS) approach. Each of these methods has its advantages and its disadvantages (Green, 1997) but the speed of the WGS method has made it the method of choice for the generation of low coverage (2x) draft sequences of many vertebrate species. A hybrid approach may also be used that combines both of these approaches.

1.6.1 Hierarchical sequencing and assembly

The hierarchical approach relies on the use of clone libraries where each clone covers a fragment of the genome, and the location of each of these fragment is known. Typically bacterial artificial chromosome (BAC) libraries of the genome of interest will be used. Each of these BACs is mapped to the genome and a set of overlapping BACs is selected that ideally cover the entire genome. These selected BACs are sequenced individually using a shotgun approach. The resulting sequence reads are assembled into a complete sequence that spans the BAC by using the overlaps present between the different reads (figure 1.10 overleaf).

Because each of the sequenced clones will be typically only 40 -200 kb in size this approach reduces the scale of the assembly problem to each individual BAC. As the location in the genome of each BAC is know if specific regions require additional finishing efforts it is easy to identify which clones to use. Problematic regions can be easily identified and targeted for re-sequencing or gap filling when required and they will not affect the sequence assembly of other regions. Because of the clone approach identification of sequencing errors will be easier as there will be no polymorphisms that may cause sequence divergence between different reads of the same region (Green, 1997).

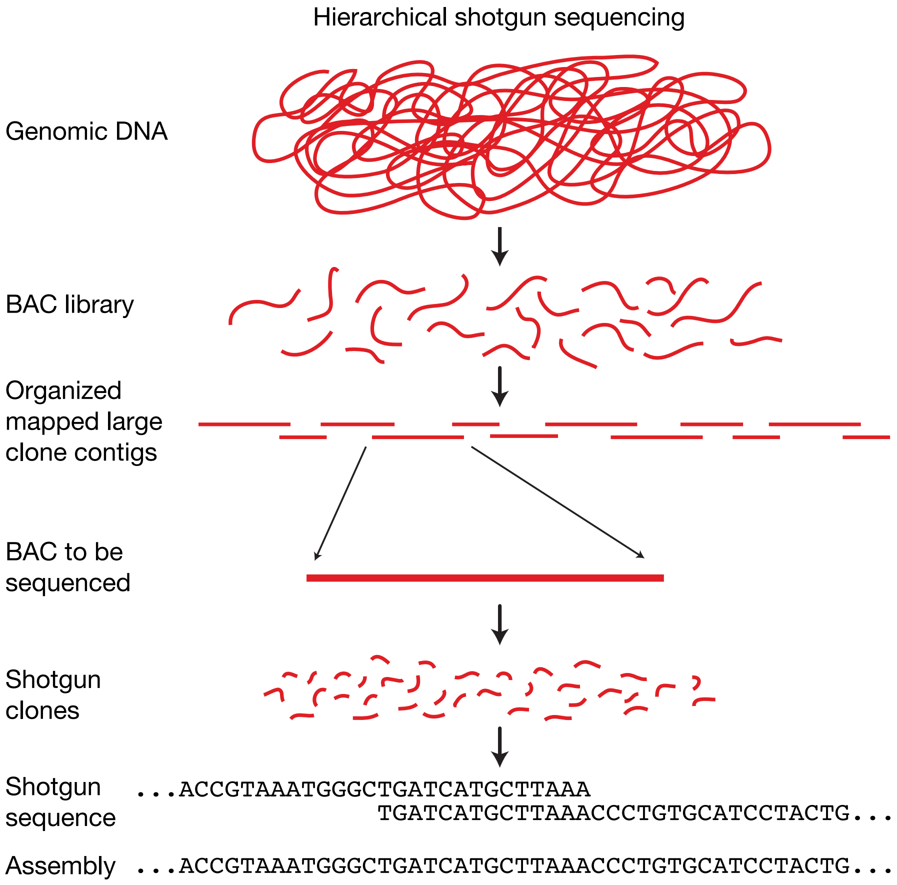

Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

Figure 1.10: Representation of the hierarchical sequencing approach used in the sequencing of the human genome. Modified from (Lander *et al.*, 2001)

This approach was the one used by the international sequencing effort in the sequencing of the human genome (Lander *et al.*, 2001).

1.6.2 Whole genome shotgun and assembly

The WGS approach does not use an initial mapping of clones to the genomic sequence. The entire genome is fragmented from the beginning and used to construct libraries of varying size. The ends of each of these resulting clones are sequenced in order to record the pairs of sequences that flank each of the regions included in the clones.

After this in the first step of the assembly all of the initial sequence reads are searched for overlaps and these overlaps are used to build contigs which will contain no gaps, although stretches of 'N's may appear in areas of sequence ambiguity (figure 1.11 A).

The WGS contigs generated in this manner are assembled into supercontigs using the relations between the contigs and the pairs of sequence ends of the initial clones (figure 1.11 B).

The advantage of the shotgun method given the current sequencing techniques is the lower cost and the speed at which an initial sequence can be obtained, as it requires no initial mapping phase.

The human genome sequencing effort from Celera Genomics was done using this approach. Most of the initial draft genomic sequence that are being generated for a sample of mammalian species also use this method. This includes the initial sequencing of the chimpanzee genome, although this project also relied on the similarity of the human and chimpanzee for the assembly phase (CSAC, 2005)

1.6.3 Hybrid approach

These two approaches are not incompatible and may be combined in different ways. Indeed, what the hierarchical approach does is divide the problem, and

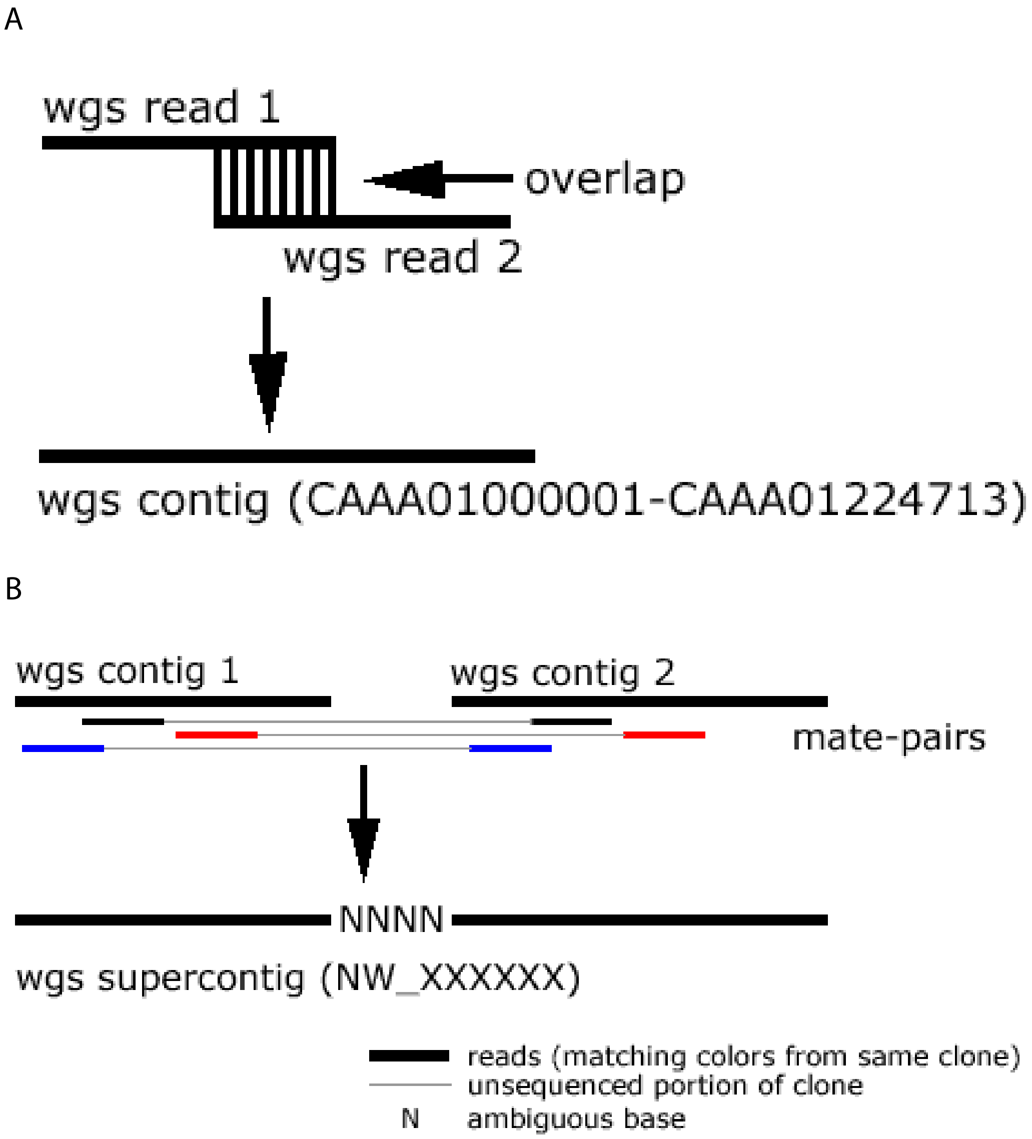

A
wgs read 1
overlap
wgs read 2
wgs contig (CAAA01000001-CAAA01224713)
B
wgs contig 1
wgs contig 2
mate-pairs
NNNN
wgs supercontig (NW_XXXXXX)
reads (matching colors from same clone)
unsequenced portion of clone
N ambiguous base

Figure 1.11: Representation of the WGS sequencing approach. A: Shows how the contigs are produced. B: Shows How the supercontigs are built from the contig sequences. Modified from NCBI (`http://www.ncbi.nlm.nih.gov/projects/genome/guide/Assembly/Assembly.shtml`)

each of the resulting fragments are sequenced using a shotgun approach. By using a more limited mapping initially the cost of this first phase can be reduced, while still having a better mapping of the resulting clones to the genome. The WGS assembly can then be based on this limited mapping. Another option is to generate the complete assembly from the WGS and later integrate available clone sequence information into this assembly. This approach was used for the sequencing of the mouse genome (Waterston *et al.*, 2002). In the macaque the result of different WGS assembly methods were combined and in addition certain regions were targeted for additional finishing (Gibbs *et al.*, 2007).

1.6.4 Genome annotation

Once the genomic sequence is assembled, the work is not completed. This stage could be compared to having a very detailed map with no labels printed upon it. A major part of the genome sequencing effort involves the accurate annotation of the different functional regions within the genome. This annotation may be done in different ways that depend on the amount of information available for the species. Particularly important is the expression data which will allow identification of the actively transcribed genes, as well as splice sites within these genes.

Accurate automatic annotation of those genomes that have a significant number of cDNA libraries available relies of the mapping of these sequences onto the genomic sequence. This requires high quality and nearly complete cDNA sequence libraries and even then it is unclear the fraction of low frequency alternative transcripts that can be recovered from them (Guigo *et al.*, 2006). Although this kind of high quality libraries are being developed for the human genome, this is unlikely to be the case for other organisms in the near future.

In these species annotation relies heavily on sequence data from closely related species, which in the case of mammals will usually be human and mouse genomes, in addition to any cDNA available for the particular species. Comparison between species will have difficulty detecting fast evolving orthologues and is likely to miss any species specific genes.

For species with little or no expression data and no closely related genome available, *ab initio* computational prediction is used, which is not the most accurate.

Because of this the accuracy of the annotation will be worse in species with little expression data particularly when there is no close species to which it may be compared (Guigo *et al.*, 2006).

There are different annotation methods available some of them use manual curation and others are completely automatic. One of the most commonly used automatic annotation pipelines are the EnsEMBL pipeline (Hubbard *et al.*, 2007). This pipeline integrates information from all the sources mentioned above in order to generate the genome annotation. It allows also for the annotation of low coverage genomes, mainly by comparison with other related species, and it also generates multiple transcripts predictions. Another of the strengths of this pipeline is the ability to keep track of annotation changes that have affected the same gene in different releases of the database.

Although the current annotation is far from perfect, the EnsEMBL pipeline was one of the top performers in a recent quality assessment of different annotation methods (Guigo *et al.*, 2006).

1.6.5 Known problems

Although the general outline of the genome sequencing methods is quite simple, as often happens when simple methods are confronted with real data unexpected situations arise. Several years after the completion of the human genome we still do not have a complete catalogue of protein coding genes in the genome. This is a very good reflection of the problems that exist even with the large amount of resources and tools developed for its sequencing and annotation. In the case of other species the situation is worse, as the amount of data available for them is not even close to the resources available for human. Many of the low level coverage mammalian sequences that are available are highly fragmented and showing large gaps and unmapped regions that have not been assigned to chromosomes. In the case of the chimpanzee and macaque genomes the situation is slightly better, mainly because of their similarity to human. This has allowed for a much more accurate comparison based annotation, however, it will still mean we will miss species specific features.

Problems affecting genome sequencing may be divided into those that affect the sequencing and assembly and those that affect the annotation of the genome once the sequencing step is complete.

Sequencing and Assembly level

Heterochromatic regions such as centromeres and telomeres cannot be sequenced due to the large amount of simple tandem repeats present in these areas, however, this is not a problem when examining gene content, as these regions are not know to contain any coding areas. On the assembly level there are difficulties assembling repeat regions, and many of them may be missed. This problem affects the WGS method in particular, although it is a problem for both when the level of identity is very high and the number of repeats large, as is the case with the ribosomal gene clusters. This problem with the assembly of repeated sequence arises because the reads that are obtained from the normal sequencing techniques are generally limited to < 1 kb in size. Regions containing many highly identical repeats cannot be discriminated when searching for read overlaps and will be collapsed into one, this problem is particularly important in the case of WGS (She *et al.*, 2004). The affected regions require additional sequencing using different techniques in order to resolve the correct number of repetitions and the sequence boundaries (Bovee *et al.*, 2008).

Currently most genomes are being sequenced as low coverage drafts using WGS and this means they are likely to be missing many recently duplicated regions (She *et al.*, 2004). Low sequence coverage has a particularly important effect on the WGS assembly, as the number of available overlapping reads will be lower, resulting in a more fragmented assembly and larger number of contigs linked by a single clone end sequence increasing the chance of assembly errors.

Annotation level

At the annotation level the difficulties mentioned above for the identification of new genes increase with the distance of the species to those high quality genomes available, as well as with the lack of expression data (Hubbard *et al.*, 2007). Some correctly predicted genes may be missing transcripts or have incorrect intron-exon boundaries. Adjacent genes may be merged into a single one and spurious introns may be predicted in order to force the reading frame conservation between an existing gene and its putative orthologue in another organism. Finally one of the major problems when using comparative genomics for gene annotation is that regardless of how well we may identify conserved genes it will be difficult to identify genes that are specific to these lineages.

1.7 Objective

In this work we will be examining the genome evolution a two different levels.

At the genomic level we will examine the differences that have accumulated between the genomes of macaque, human and chimpanzee since they diverged from their common ancestor. At this level we will focus on new gene formation.

At the gene level by examining changes in the intron-exon structure of recently formed paralogous genes in *A. thaliana*.

In the first part of our study (chapter 3 on page 55) we implement an automated pipeline capable of building a set of synteny blocks between each of the genome pairs and classifying all protein-coding genes into different groups based on sequence similarity and the synteny information obtained from the blocks.

Once this classification is complete we will use the information obtained from each of these pairwise comparisons in order to identify lineage specific differences in each of the two hominoids examined such as gene extinction/creations and translocations. We will also identify those genes that are conserved as 1:1:1 orthologues in all three species.

In the second part (chapter 4 on page 107) we examine in detail those protein-coding genes that were identified as lineage specific in either the human or chimpanzee lineage. Within these genes we searched for likely cases of *de novo* gene creation by the exaptation of an area of non-coding DNA.

In the third part (chapter 5 on page 133) we examine another of the processes that may affect the gene content of an organism, this is alternative splicing with subsequence subfunctionalization. We searched for cases in which a pair of genes in one of the two organisms may have originated by this process.

In the fourth part of this work (chapter 6 on page 171) the mechanism of intron gain and loss is examined. In this case by using data from the plant *Arabidopsis thaliana*.

Finally in the last part of this work (7 on page 197) we present a summary of the obtained results, as well as some of the conclusions and ideas that may be derived from them.

Chapter 2

Materials and Methods

2.1 General Methods

2.1.1 Data

All the sequence and location data were obtained from the EnsEMBL database (Hubbard *et al.*, 2007), the version used was release 46. EnsEMBL was chosen because it is capable of predicting multiple transcripts for a gene and was one of the best performers in the EGASP tests (Guigo *et al.*, 2006). Although manually curated data is considered the golden standard, the data available with this quality are much smaller (not even half of the complete set of human genes have been curated manually) and does not allow for the kind of analysis we intended to perform.

The EST data was downloaded from `ftp.ebi.ac.uk` on 26 March 2007.

2.1.2 Data extraction and analysis

The analyses described were done using scripts written in PERL with the help of the BioPerl (`http://www.bioperl.org`) modules and EnsEMBL API (`http://www.ensembl.org`), R, which was used for statistical analyses and MySQL that was used for data organization and storing.

2.1.3 BLAST searches

In all cases in which it is not otherwise specified, BLAST searches were performed using NCBI BLAST with an E-value cutoff of $1e-4$ and a fixed database size of $1e9$. This fixed database size was used in order to make e-values obtained from different database searches comparable.

BLAST searches against the genome trace data were performed using the NCBI BLAST website (`www.ncbi.nlm.nih.gov/blast/`).

2.1.4 Sequence masking

Masking of the primate genomic sequences used for alignments was done using RepeatMasker (`www.repeatmasker.org/`).

2.1.5 Alignments

The alignment program t_coffee (Notredame *et al.*, 2000) with default parameters was used for all protein level alignments unless otherwise stated.

In the case of nucleotide level alignments of the genomic sequence from the three primate species MultiPipMaker was used (Schwartz *et al.*, 2000).

2.1.6 EnsEMBL Assembly evaluation

Small gaps within the assembled chromosomes were identified by the presence of long stretches of "N"s that are inserted by EnsEMBL in those regions where a missing or incorrectly sequenced area is present.

2.2 Orthology

Orthology (ortho=exact, logos = ratio, proportion) is the relation between two genes that originated through a speciation event as opposed to any other gene formation mechanism. This relation is defined only with respect to the phylogeny of the genes, independently of their function. However it is assumed that in most cases orthologous genes in different organisms will perform the same functions as the ancestral gene from which they derived.

There have been many methods that have been used in order to identify groups of orthologous genes between different organisms (Tatusov *et al.* 1997, Huynen and Bork 1998, Alexeyenko *et al.* 2006, Wapinski *et al.* 2007), most of them can be grouped within three main types:

- **Hit-clustering methods:** These methods group genes using some measure of similarity, usually the e-value from a Smith-Waterman type search (Snel *et al.*, 1999). They have the disadvantage of not resolving the evolutionary relationships, and have trouble with the many-to-many relations which exist in real data due to gene duplications. In some cases if only reciprocal best hits (RBHs) are used they ignore paralogy completely. One of the biggest problem in the RBH methods is the inclusion of new species, as with a larger number of species the number of RBHs shared between all of them decreases. An example of this problem can be seen in the antifreeze proteins of cod and notothenioid fish which have nearly identical sequence but different origins (Chen *et al.*, 1997). This method works well in bacterial genomes, however in the case of eukaryotes it ignores the fact that many eukaryotic genes can have multiple splice variants each of them with a different RBH in the other species.

- **Tree based methods:** For each family a tree is built, and the orthology relations are determined by comparing it to the species tree. By using reconciled trees duplication and loss events can be identified. The main problem with these methods is the uncertainty of the tree topology, both for species and gene trees. Trees that deviate from expected for any reason are likely to cause the incorrect prediction of duplication or loss events (Huerta-Cepas *et al.*, 2007).

- Synteny based methods: These cannot be applied to distantly related species where many rearrangements have occurred since their divergence time (Huynen and Bork, 1998). An example of synteny based method has been used by Goodstadt to create synteny maps of human vs dog (*Canis familiaris* and human vs. opossum (*Monodelphis domestica*) (Goodstadt and Ponting 2006, Goodstadt *et al.* 2007).

These methods can also be used in combined approaches, in order to take advantage of the strengths of each one of them. Indeed in many cases the initial step of most modern orthology assignment methods will be a Smith-Waterman search, that will then be followed by the addition of phylogeny and/or synteny based approach.

Problems with the identification of orthologues that will affect all methods to a greater or lesser degree include sequence divergence in fast evolving genes, non-orthologous gene displacement, and changes in gene content by the previously mentioned mechanisms of gene duplication, loss, horizontal gene transfer,or gene fusion/fission. Because of these problems, and those caused by the rapid turnover, and imperfect duplication of genes, that may generate partial and chimeric duplications, it might not always be possible to determine which are the real orthologues of two genes from distant species, and it has been suggested that a classification into complete partial or chimeric orthologues may be more adequate (Katju and Lynch, 2006).

In the case of the primate genomes, because of the short divergence times since the separation of their lineages from their common ancestor, we chose a synteny based approach as the best method for identifying orthologues between the tree species.

Chapter 3

Gene content differences between human and chimpanzee

3.1 Introduction

The differences between the primates, as between other organisms, are likely to originate from the combination of several factors such as differences in gene content, in gene copy number (Kehrer-Sawatzki and Cooper, 2007a), sequence differences between orthologues (CSAC, 2005), and differences in orthologue regulation in the different species (Asthana *et al.*, 2007). With the availability of the complete human (*Homo sapiens*) genome sequence (Lander *et al.* 2001, IHGSC 2004), as well as those for the chimpanzee (*Pan troglodytes*) and macaque (*Macaca mulatta*), which were recently completed (CSAC 2005, Gibbs *et al.* 2007) it should be possible to determine which regions and which genes are conserved between these species as well as which are unique to one genome, and determine how much each of these factors contribute to the phenotypic differences.

Although this may seem easy enough, the varying results obtained from different studies (CSAC 2005, Gibbs *et al.* 2007, Demuth *et al.* 2006, Blomme *et al.* 2006), show it is not a trivial matter. When examining the problem in detail we can see several problems arising, such as assigning orthology correctly, the completeness of the genomic sequence, and the reliability of the genome annotation. All of these will influence our ability to determine which genes are unique to one species, which ones are shared, and in what cases there is a difference in gene copy number between species.

Orthology assignment is a complex problem, as can be gathered from the amount of resources invested in generating orthologue sets with various methods since the first major effort on automatic orthology assignment (Tatusov *et al.*, 1997) and the differences in the sets of orthologues returned by each one of them (Alexeyenko *et al.*, 2006). Nevertheless it is crucial to assign orthology correctly in order to know which genes are conserved between different genomes.

The quality of the finished human genome (HG) is very good (IHGSC, 2004), however the cost of producing a finished state genome is very high, and many available genomes, such as that of chimpanzee, are in a Working Draft (WD) state only and there is currently no intention of producing a finished version (Taudien *et al.*, 2006). The percentage of the genome missing varies, as well as its coverage – the chimpanzee genome covers 94% of the species' genome (CSAC, 2005), and the macaque genome covers 98% (Gibbs *et al.*, 2007). Although this will give us a very good indication of the overall gene content, we could still be missing many genes. Another source of organism differences and also of assembly problems are segmental duplications (Wooding and Jorde, 2006). If the duplicated areas have high identity, as would be the case in the most recent ones, they are likely to be missing completely in genomes sequenced only by the Whole Genome Shotgun (WGS) method, and if present, they may have been collapsed into one. This means that any genes duplicated within these regions are likely to appear as unique copies if they are present at all (She *et al.*, 2004).

The difference in sequence content between the human and the chimpanzee draft genome caused by areas contained in segmental duplications greater than 20 kb was estimated to be 70 Mb with 177 human specific and 94 chimpanzee specific genes annotated within them (Cheng *et al.*, 2005), however smaller duplications are likely to still be missing from this estimate. What we have learned from the HG project has confirmed these issues, as most of the larger missing areas are located in heterochromatic regions and in or close to areas of segmental duplication with a high identity (Schmutz *et al.*, 2004). Because of this, we cannot assume the absence of a gene in the genome annotation means the gene is not present in the organism, although the likelihood of this being the case increases with the coverage of the genome. The results of several recent studies of gene content in vertebrates (Demuth *et al.* 2006, Blomme *et al.* 2006), where those genomes with a greater number of gains and fewer losses seem to be those for which the genome coverage and assembly quality are higher, hint at this problem (figures 3.1 overleaf and 3.2 on page 59). The genome version used was actually noted as an important factor affecting the perceived gene gain and loss in one of these studies (Demuth *et al.*, 2006).

Most of the recently sequence genomes are low coverage WD genomes that have been sequenced using the WGS method, this means they are likely to be affected by some of these problems.

Finally, we also have to face the annotation problems arising from the difficulty of accurately predicting gene structure using computational methods. Even in the human genome, the most extensively studied, there is no complete set of protein coding genes available. Although very few genes, if any, seem to be missing from the computational predictions, the exact genomic structure and alternative splicing patterns are estimated to be correct in only 50% of them (Guigo *et al.*, 2006). With a huge fraction of mammalian genes apparently possessing alternatively spliced variants (Kim *et al.*, 2007), this is a big problem. The situation in other genomes that have not been studied in so much detail is unlikely to be better, and many lineage-specific genes may remain un-annotated or un-sequenced. This could account in part for the surprising fact that the HG seems to show more gains, fewer losses and more unique genes than any other organism with which it has been compared, even though the human genome is

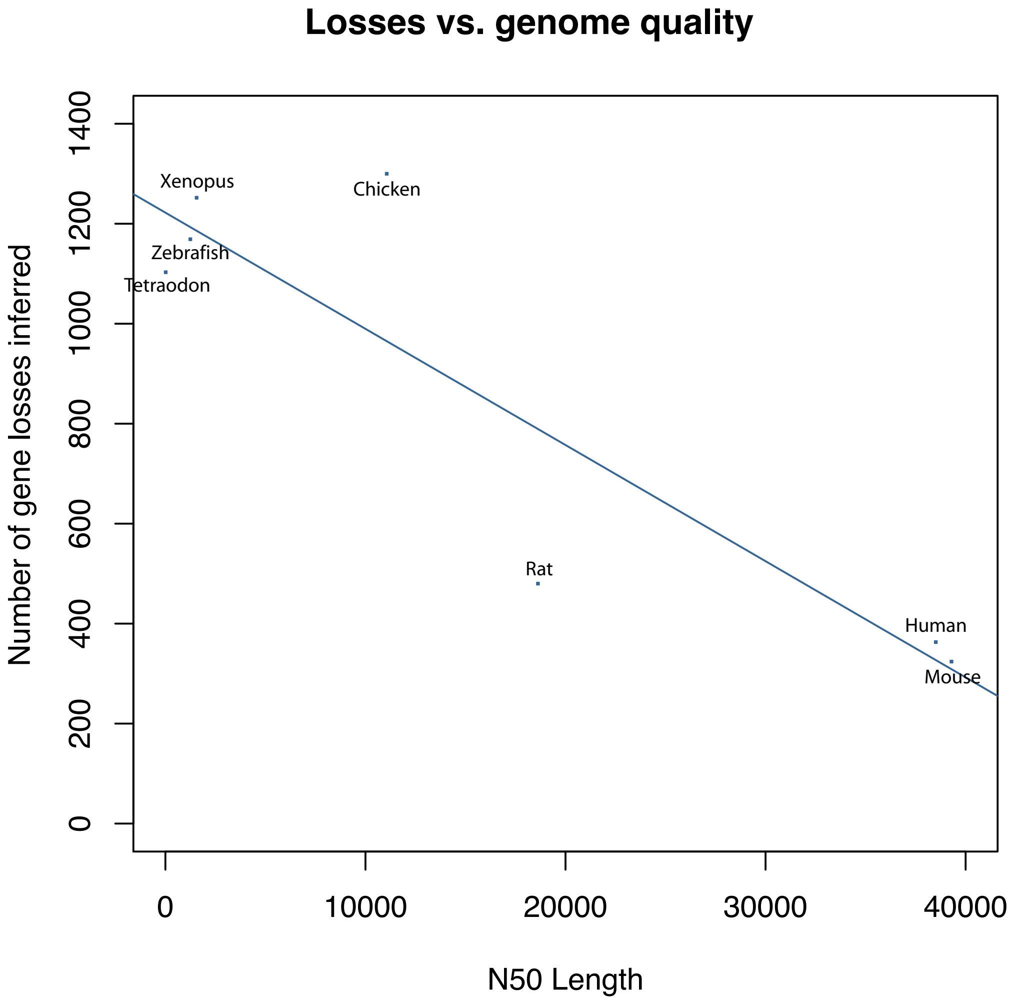


Figure 3.1: Comparison of the genome quality, measured by the N50 length (L_{N50}), with number of gene losses found by Blomme *et al.* (2006). L_{N50} is defined in such a way that half of the nucleotides of the genome are located in continuous contigs or scaffolds of at least the L_{N50} size. The L_{N50} values used are from the currently available genomes for each of the species, which means that in some cases the L_{N50} may have increased since the time of the studies.

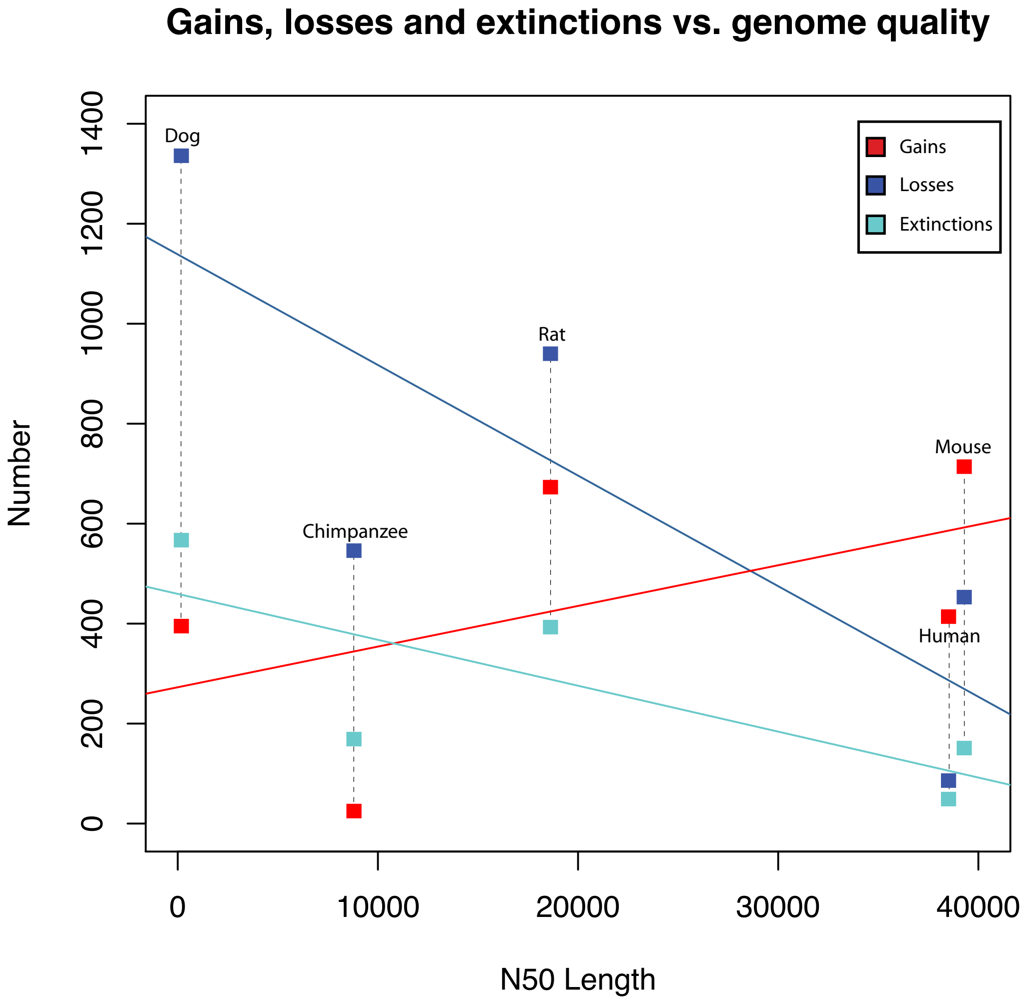


Figure 3.2: Comparison of the genome quality, measured by the N50 length (L_{N50}), with number of gene gains (red), losses (blue) and extinctions (cyan) found by Demuth *et al.* (2006). Extinctions are a subset of losses defined as the loss of the last gene in a family. The L_{N50} values used are from the currently available genomes for each of the species and may have increased since the time of the study. This is the case for the chimpanzee genome

smaller than the other hominoids for which the C-value is known.

Although this situation seems to indicate that we cannot extract any relevant information from the available genomic data, this is not the case. However, the analysis of gene gain and loss in these genomes must be conducted with extreme rigor controlling and checking for sequencing gaps and annotation errors.

Using one-to one orthologues we identified regions of conserved gene order between the three sequenced primates. Within these regions of conserved gene order we searched for difference in gene presence and absence. This synteny framework encompasses an expected location for each gene. We exploit this to account for sequencing and annotation artifacts in those cases in which a gene appears to be present in one of the species but not in the other.

3.2 Materials and Methods

3.2.1 Data

The sequence data were obtained from EnsEMBL, see section 2.1 on page 51.

The N50 length value used as a measure of the quality of each genome was obtained from NCBI (`http://www.ncbi.nlm.nih.gov/Genomes`) except for the numbers corresponding to *Canis familiaris* where it was obtained from Lindblad-Toh *et al.* (2005), *Xenopus tropicalis* where the information was obtained from the Doe Joint Genome Institute (`http://genome.jgi-psf.org/Xentr4/Xentr4.info.html`) and *Tetraodon nigroviridis* where the information was obtained from Genoscope (`http://www.genoscope.cns.fr/externe/tetranew`).

C-values were obtained from the Animal Genome Size Database (Gregory, T.R. 2008 `http://www.genomesize.com`).

3.2.2 Sequence similarity search and grouping

For each pair of species (*H. sapiens* - *P. troglodytes*, *H. sapiens* - *M. mulatta*, and *P. troglodytes* - *M. mulatta*) all the genes labeled by EnsEMBL as protein coding, V segment or C segment were extracted from the database. These are the genes that have a protein product annotated. A BLAST database was built with the proteins they encoded, including all alternative transcripts if any. We removed 369 *H. sapiens* genes belonging to the haplotypes *c22_H2* (1), *c5_H2* (18), *c6_COX* (182) and *c6_QBL* (168), as they are redundant.

An all-against-all BLASTp search was done for these proteins using a cut-off E-value of $1e^{-4}$ and a fixed database size of $1e^{9}$ to ensure e-values from the different database searches were comparable. For brevity and clarity whenever we refer to a BLAST hit in which the product of gene A hits the product of gene B we will call it a hit from gene A to gene B.

We considered all BLASTp hits with an e-value within a range of 10^{3} from the e-value of the top hit to be equally good hits. This approach controls for some of the coarseness of e-values, and ensures that we only place confidence in top hits (or groups of top hits) that are clearly distinguished from other similar genes when identifying orthologues.

These Reciprocal Top Hits (RTHs) between the proteins of each pair of species were extracted, and their corresponding genes identified. Because we are including all transcripts from each gene the different products of one gene may have top hits to more than one gene in the genome under comparison. The genes were split into groups by grouping those genes that hit each other as RTHs into the same group. The resulting groups were formed by closely related genes where every member of the group hit at least one other gene in the group as a RTH.

The Initial Orthologues (IOs) were selected as a subset of these groups using a much stricter criterion. They were defined as gene pairs whose products only hit each other within the search range (*i.e.* one-to-one reciprocal best hits). In order to avoid close paralogues we did not include in this IO set any genes with proteins that hit another protein encoded by a different gene within this threshold of $1e^{3}$ from the e-value of the best hit.

3.2.3 Synteny block construction

We used the IOs to build non-overlapping synteny blocks in which each pair of IOs were separated by 10 genes or less from the next pair in both organisms (figure 3.3 overleaf). Small inversions within the blocks were kept (spanning fewer than 10 genes), however if they were included in the expected location of a gene A (section 3.2.4 below) this expected location was considered to be ambiguous.

The resulting blocks were searched for overlaps. If an overlap was caused by genes from different blocks being interleaved at the end of their respective blocks, the area involved was removed from both blocks. If the cause was a small block included in a larger one the large one was split and the small one was kept resulting in three non overlapping blocks.

3.2.4 Defining the expected location of a gene

We define the expected location of a gene as a conceptual ten gene window on either side of the location of the gene of interest. Because the location of the orthologue of the gene of interest is unknown this distance is projected on the other genome through the IOs on either side of the gene of interest. In this way we obtain the equivalent region of the other genome within which we expect to find the orthologue of the gene of interest (figure 3.4 overleaf). If the area contains a small inversion this expected location is considered ambiguous, as the genes that are closer to the gene 1 in genome A may have an inverted orientation along the chromosome, and this means the gene could easily be outside of the 10 gene threshold we assigned, but still be present.

3.2.5 Gene classification

Using the RTHs for each of the genes as well as the synteny information the genes were classified into the following groups.

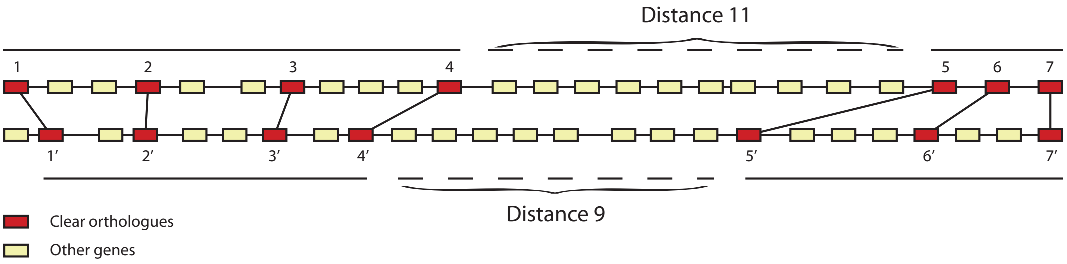


Figure 3.3: Diagram showing the method used to build the synteny blocks. The IOs are marked in red. The blocks were built by joining pairs of orthologues that were separated by fewer than 10 genes in both genomes. In the figure we can see two blocks built because there are more than 10 genes separating two contiguous orthologues in one of the two genomes.

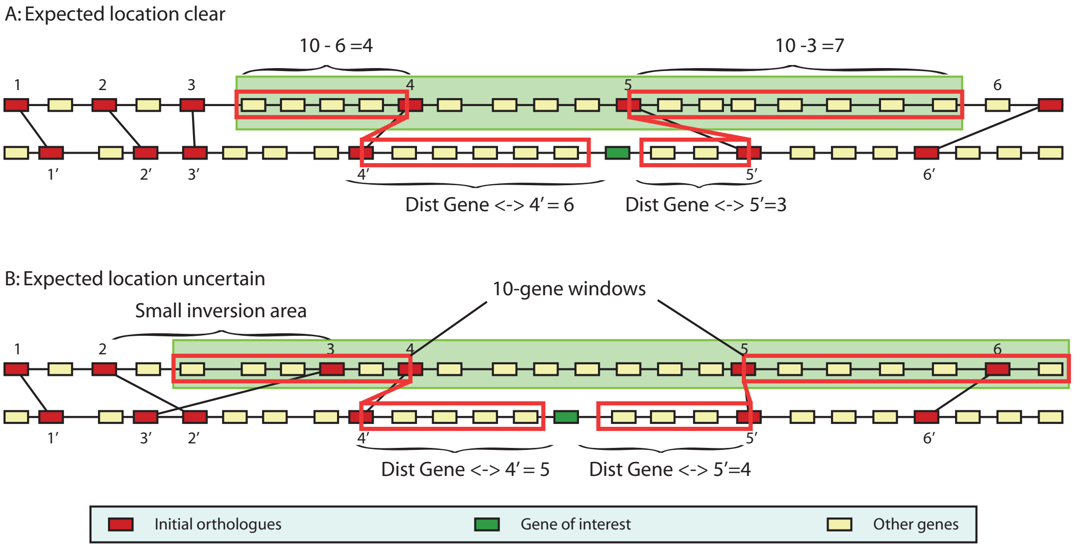


Figure 3.4: Diagram showing the method used to calculate the expected location of a gene (shaded in green). The red box indicates the conceptual 10-gene windows projected from the location of the gene of interest across the genomes via the IOs. A) The expected location is clear. B) The expected location is marked as uncertain because it overlaps an area containing a small inversion.

1. **Orthologues**: This includes the IOs, excluding those pairs in which both members are located outside of the synteny blocks, and also those pairs of genes that only hit each other as *reciprocal* top hits and are at each others expected location, but were not classified as IOs because of *non-reciprocal* top hits to another gene .

2. **Tandem duplications**: Groups of more than two genes each of them with only RTHs at their expected location no top hits outside the expected.

3. **Dispersed duplications**: Groups of more than two genes which hit each other as RTHs where at least one gene has a RTH outside of its expected location on the other genome.

4. **Transposition**: Pairs of genes which only hit each other as RTHs but they are not at each others expected location.

5. **Single genes**: These are genes with no hit in the genome of the other species within the initial BLAST threshold of $1e^{-4}$.

6. **Uncertain**: These genes have no RTHs at the expected location, or they include small inversion areas at their expected location. Those genes belonging to a group where one of the genes was classified as uncertain were also classified as uncertain.

7. **Excluded**: Groups of genes in which all members of the group are outside of the synteny blocks.

3.2.6 Evaluation of the plausibility extinction/creation candidates

A strict criteria was used to identify and remove possible false positives within the extinction/creation candidates. A diagram of the process can be seen in figure 3.5 on page 67.

Presence as a pseudogene in the other species

All genes with no BLASTp hit in the initial search were compared to all the pseudogenes annotated in EnsEMBL.

Sequence similarity searches within the "expected location"

A Smith-Waterman search using the program ssearch was used to at the nucleotide level with a cut-off of $1e^{-4}$.

At the protein level a search against all the proteins annotated at the expected location was performed by aligning each of the alternative proteins of the candidate gene with all those proteins at the expected location. This alignment was done using t_coffee and evaluated with infoalign from the EMBOSS package.

A BLASTp was also performed at the protein level against the proteins annotated at the expected location using the same settings as the initial BLAST search $1e^{-4}$ and a set database size of 10^9 but without masking.

BLAT was used in translated mode to compare the annotated proteins of each gene with the nucleotide sequence at the expected location in order to identify any hits at the expected location that may have been missed by the annotation. Those hits that had no in-frame stop codon, and showed an identity of 90% or more in all of the exons were discarded as candidates for lineage specific genes due to possible annotation errors where a gene had been missed by the annotation but was likely to be present in the genome.

Evidence of the presence of a similar gene in the other genome

In order to identify cases where the gene may have been relocated, we also checked if there were any homologues to these genes annotated by EnsEMBL in their pipeline.

Supporting evidence for the gene veracity

A tBLASTx against the EST database was done to determine if the extinction/creation candidate was indeed expressed. If one of its transcripts had at least two different EST hits from the same species with an identity of 100% over a stretch of 33 amino acids, it was considered to be expressed, and therefore a real gene.

We checked for the presence of any Vega/Havana annotation for a gene, or if the gene had an EnsEMBL status of KNOWN.

Genome sequence integrity

The expected location for the gene was checked for gaps in the assembly that could be big enough to span the gene completely.

3.2.7 GO term distribution

GO analysis was only done on the human gene groups, as 92% of the chimpanzee proteins and 95% of the macaque proteins annotated lack any GO annotation and any result obtained from their analysis could be caused by a bias in the fraction of GO annotated genes. The GOslim set of terms used was the generic GOslim from the Gene Ontology website (`www.geneontology.org`).

When determining if the distribution of GO terms in a group of genes was different from expected, we built the expected distribution for a group of that size by generating 10,000 random samples with the same number of genes from the genome of that species. The p-value was calculated from this expected distribution and corrected for multiple testing using both the Bonferroni correction, and the Benjamini-Hochberg false discovery rate (Hochberg and Benjamini, 1990). If a result is mentioned as significant, this means it remained significant after applying Bonferroni correction, which is the strictest of the two, unless otherwise noted.

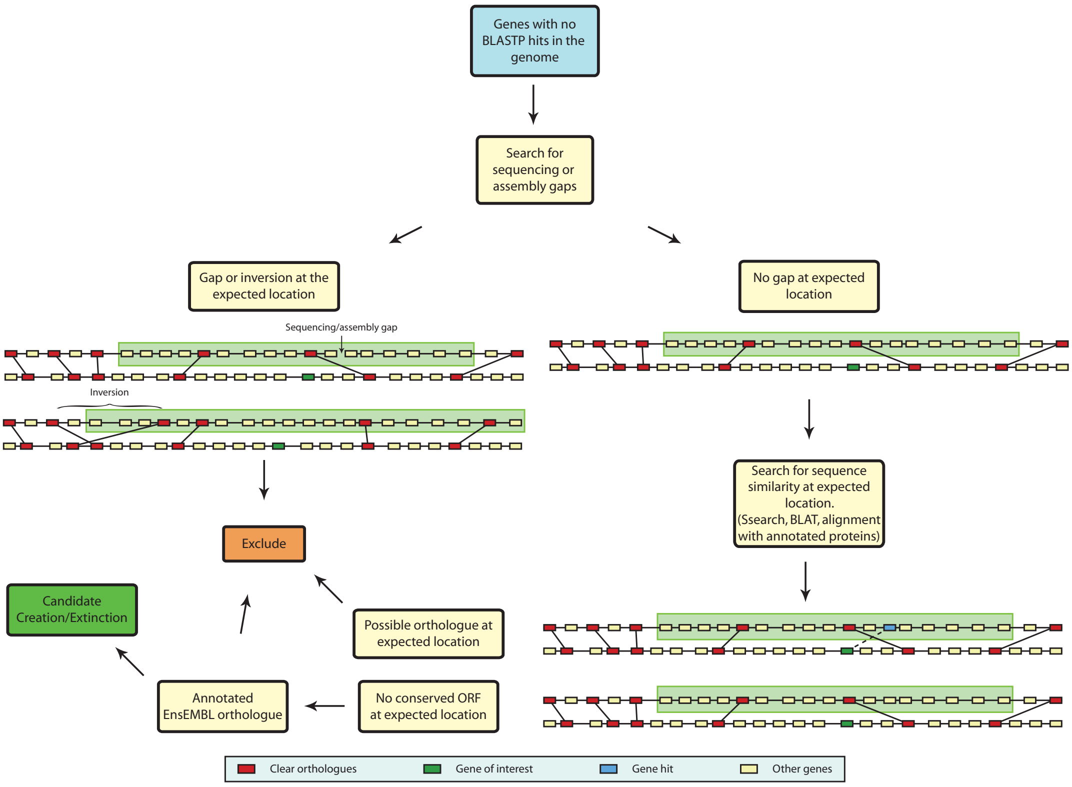

Genes with no BLASTP hits in the genome
Search for sequencing or assembly gaps
Gap or inversion at the expected location
No gap at expected location
Sequencing/assembly gap
Inversion
Search for sequence similarity at expected location. (Ssearch, BLAT, alignment with annotated proteins)
Exclude
Candidate Creation/Extinction
Possible orthologue at expected location
Annotated EnsEMBL orthologue
No conserved ORF at expected location
Clear orthologues
Gene of interest
Gene hit
Other genes

Figure 3.5: Flowchart showing the process used to eliminate potential false positives from the extinction/creation candidates. The expected location of the gene of interest in the genome of the other species is indicated by a green box

To avoid bias caused by the different sizes of the duplicated gene groups, GO analyses that involved these genes were repeated choosing groups of genes randomly instead of individual genes, and only those categories that remained significant were reported.

3.3 Results

The number of genes annotated by EnsEMBL as protein coding in human (22,568 genes excluding haplotypes), chimpanzee (20,572) and macaque (21,944) are quite different. Some of these genes are shared between the three species in single copy, while others may have a different number of copies in different species or be specific to one of the species because of *de novo* acquisition or inactivation in the other (Wang *et al.* 2006, Gilad *et al.* 2005). Shared genes might have remained in the same genomic location since the divergence from the common ancestor or may have relocated to different areas of the genome. In this work we examine these events by comparing the genome sequences available for these three primates.

We built a set of orthologues that are conserved between each pair of species, and used them to identify regions of conserved synteny. After this we classified the genes of each species into different groups according to their presence in the other species, similarity to other genes and location in the genome (see methods 3.2.5 on page 62). The groups were analyzed in search for differences in the type or genomic distribution of the genes included in each of them, and we combined the results of these pairwise comparisons to determine if any of the differences are lineage specific. An overview of the complete analysis pipeline for each pair of species can be seen in figure 3.6 overleaf.

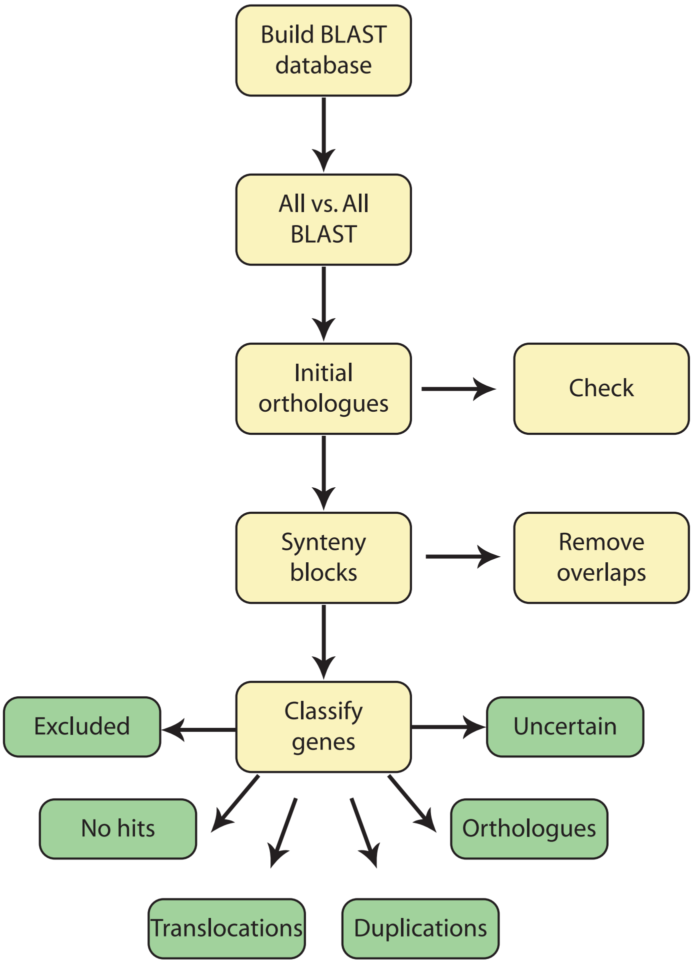


Figure 3.6: Overview of the analysis pipeline. The results of the all-against-all BLASTp search were used to obtain a group of Initial Orthologues (IO). These we checked for any inconsistencies and used to build the synteny blocks. Overlapping areas were resolved by splitting the blocks or removing the uncertain areas. The combined information from the BLAST search and the synteny blocks was used to classify the remaining genes. These are classified into six groups: Excluded –with no synteny information, Uncertain –no RTHs, No hits, Orthologues –one-to-one orthologues, Transpositions –only one RTH which is out of the expected location but within the synteny blocks, and Duplications –groups with more than two genes in one of the two species compared.

3.3.1 Orthologue detection

We inferred orthology of primate genes under strict and rigorous criteria to limit artifactual effects on the analysis of gene gain and loss. We conducted an all-against-all BLASTp search of all proteins annotated for human, chimpanzee and macaque within the EnsEMBL database (including all splice variants). We considered the best hits as not simply the first hit listed in the BLASTp output file, but as all hits that had an e-value within a rage of 10^3 of the lowest e-value. This avoids the pitfall of accepting the top hit as the likely orthologue when there are other equally, or almost equally, good hits. In each species pair comparison, gene pairs that share reciprocal best hits (RBHs) only with each other and had no nonreciprocal best hits were considered unambiguous orthologues (”Initial Orthologues” in tables 3.2 on page 75 and 3.3 on page 76). Genes that shared RBHs with more than one other gene were considered together as an RBH group, which may in reality be co-orthologues or highly similar paralogues (see Methods in section 3.2.2 on page 60).

3.3.2 A synteny block framework for the investigation of gene gain and loss

For each species pair we identified a set of genes that are clearly present in a single copy in both organisms (see section 3.2 on page 60). These initial orthologues (IOs) were used to build a set of synteny blocks in which the gene order is conserved between the genomes of both species (table 3.1 on page 72).

The larger syntenic areas, with 20 genes or more, that are conserved between human and chimpanzee, between human and macaque, and those common to the three species are shown in figure 3.7 overleaf. Most of the genome in all three species is contained in these regions.

When comparing human and chimpanzee the largest number of synteny breaks was found in chromosomes 1, 6, 16, 17 and 19 with 9 or more breaks in each.

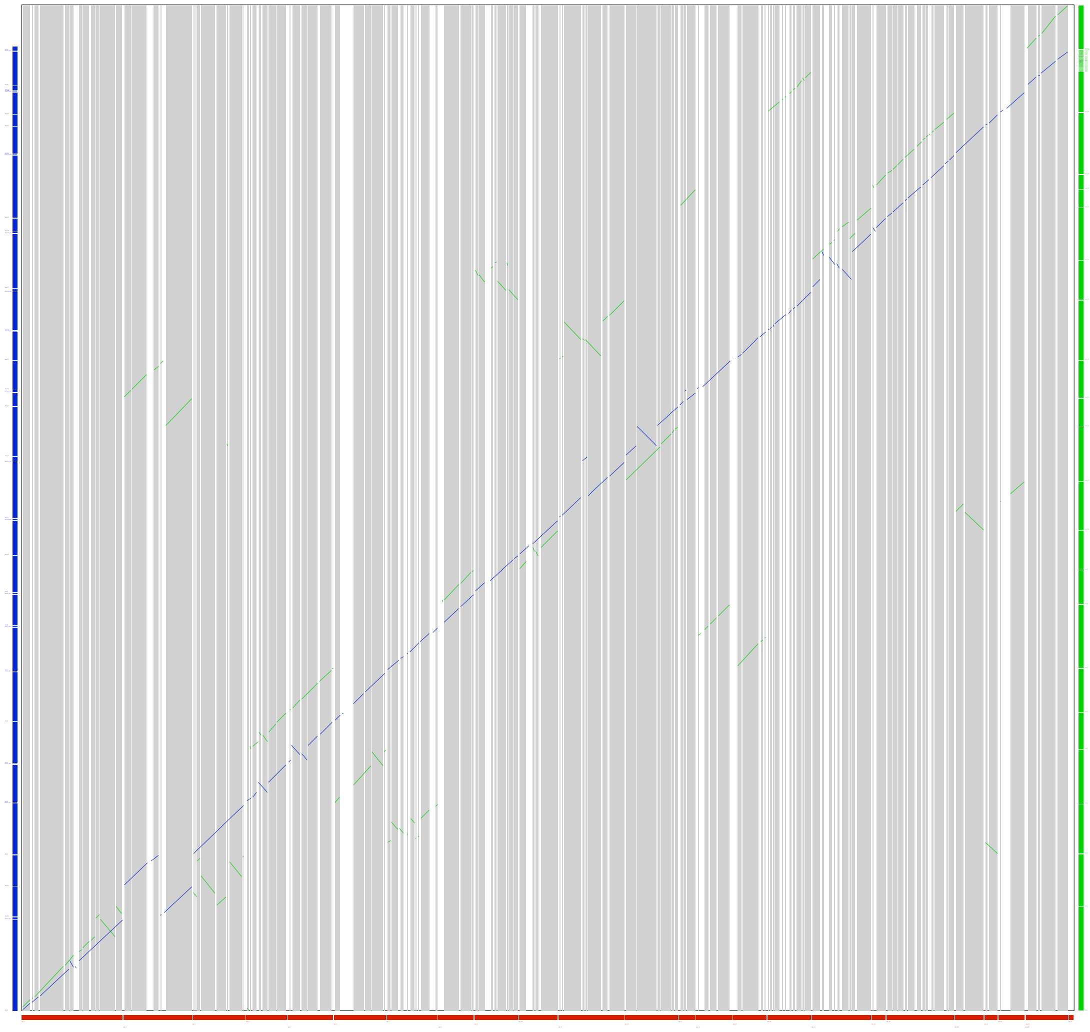


Figure 3.7: Dotplot of the conserved synteny blocks between the three primate species. The three coloured axes represent the chromosomes and/or contigs in the three genomes: human (red), chimpanzee (blue) and macaque (green). The diagonal lines represent regions where at least 20 genes show conserved synteny between human and chimpanzee (blue) or human and macaque (green). The grey shading shows the overlap of these regions where the synteny is conserved between all three species.

Table 3.1: Synteny blocks summary.

Species	Blocks	Genes included[a]	Sequence included[b]
H. sapiens	138	21195 (94%)	2.81 Gb of 3.1 Gb (91%)
P. troglodytes		19634 (95%)	2.84 Gb of 3.4 Gb (85%)
H. sapiens	213	20415 (90%)	2.68 Gb of 3.1 Gb (87%)
M. mulatta		20078 (91%)	2.68 Gb of 3.1 Gb (87%)
P. troglodytes	229	19236 (94%)	2.74 Gb of 3.4 Gb (82%)
M. mulatta		20183 (92%)	2.70 Gb of 3.1 Gb (86%)

[a] Number of genes included in the synteny blocks and percentage of the total protein coding genes they represent

[b] Number of nucleotides and percentage of the total genome size included in the synteny blocks. This is the fraction of the sum of the lengths of all non-redundant top-level regions (Golden Path Length)

Synteny conservation between human and macaque is also good, but lower than between human and chimpanzee. There are 11 human chromosomes with 9 or more synteny breaks when compared to macaque with the largest numbers in chromosomes 1, 7, 16 and 19 with 14 or more breaks each. The fragmentation of the macaque assembly, and the greater phylogenetic distance between these two species explain this lower continuity of the synteny blocks.

Although the syntenic regions between chimpanzee and macaque cover a fraction of the macaque genome similar to the comparison with human, the fragmentation of the blocks is greater. There are 10 chimpanzee chromosomes with 9 or more synteny breaks, one less than in the human comparison. The ones with the largest number of breaks are also chromosomes 1, 7, 16 and 19 with 13 or more breaks each. The reason for this greater discontinuity is probably the number of genes in both of these organisms that are located in regions which are not yet assembled into the final chromosomes, but still appear as scaffolds, or as belonging to a certain chromosome, but in an uncertain position (EnsEMBL chromosome random annotations)

The number of genes belonging to orthologue pairs in chimpanzee chromosome 19, and human chromosome 19 was significantly lower than expected if the distribution of these genes among the chromosomes was random. In the case of the comparison between human and chimpanzee, this result is significant when using the Benjamini & Hochberg false discovery rate correction for multiple testing (Benjamini and Hochberg, 1995) with $P \leq 0.04$, but not when applying the Bonferroni correction ($P \leq 0.11$). An explanation for this can arise from the observation that the average rate of synonymous substitution of those genes located on human chromosome 19 is extremely high compared to other chromosomes (Castresana, 2002). This may indicate a more rapid divergence between orthologues that would hamper orthologue identification. Another factor that could contribute to this observation is the large number of segmental duplications and tandemly arranged gene families in this chromosome (Grimwood *et al.*, 2004). This includes a group of DNA binding proteins that constitutes one of the strongest clustering of gene functions in the genome (Castresana *et al.*, 2004). The excess of genes classified as tandem duplicates in this chromosome for both of these species when compared to macaque seems to indicate these particular properties of chromosome 19 originated before the common ancestor of the three species, but after the divergence from the mouse lineage (Castresana, 2002). The high number of synteny gaps in this chromosome between human and chimpanzee compared to other chromosomes may be another consequence of these properties. A large number of duplicated regions could also explain the cases of chromosomes 7 and 16, which have some of the highest fractions of intra-chromosomal segmental duplications in the human genome (Lander *et al.*, 2001).

There are also major rearrangements that have occurred between human and chimpanzee in chromosomes 1, 16 and 17, and between chromosomes 1 and 7 between the Hominoidea and Cercopithecoidea lineages that could account for some of these large numbers of breaks (Gibbs *et al.*, 2007).

3.4 Gene classification

Because gene order within a synteny block is conserved, orthologues of genes present in one genome and included in a synteny block would be expected to maintain the same relative location in the other genome if present unless there has been a rearrangement. Based on this assumption we classified the rest of the genes in the two species' genomes into five different classes (table 3.2 overleaf).

The same classification, but excluding those genes in which one of the members of the group is outside of the synteny blocks, can be seen in table 3.3 on page 76. This may represent in a better way the real situation, as in many cases genes that are not included in a synteny block are located in areas of the genome that are not completely assembled in one of the species. We excluded the cases in which one of the members of the group is out of the synteny blocks, because if this is the case the group will be classified as a transposition (of one gene or a dispersed duplication if it includes more than 2 genes) regardless of whether this is really the case.

3.4.1 Genes that are conserved in single copy (Orthologues)

We identified those genes that were single copy orthologues in both genomes for each of the genome pair comparisons (table 3.2 overleaf). Within these genes 13274 were 1:1:1 orthologues .

The GO term distribution of the genes belonging to orthologue pairs between human and the other two primates that show the terms *nucleus*, *protein transport*, *cell cycle* and *transcription factor activity* were significantly over-represented, whilst there was a significantly lower number of genes with the terms *calcium ion binding* and *chromosome*.

Lopez-Bigas *et al.* (2008) showed that in mammalian genomes regulatory genes diverge more quickly and genes in core processes tend to be conserved.

Table 3.2: Gene classification for each of the species pair comparisons, as well as the number of gene groups (Gr) in each category.

			H. sapiens			***H. sapiens***			***P. troglodytes***		
			P. troglodytes			***M. mulatta***			***M. mulatta***		
Gene type[a]			**Hs**	**Pt**	**Gr**	**Hs**	**Mm**	**Gr**	**Pt**	**Mm**	**Gr**
Orthologues	1	Initial	15915			13310			13203		
		Synteny	364			393			312		
	4	Transposed[b]	123			161			172		
		Total	16402			13864			13687		
Duplications[c]	2	Tandem	405	328	178	435	375	223	317	308	172
	3	Dispersed	2774	2486	1027	2833	2989	1173	2511	2852	1102
Single genes	5		644	136		1748	347		1358	487	
Uncertain	6	1298	649	71	1875	2960	94	1572	3204	81	
Excluded	7	1045	571	31	1638	1234	27	889	1168	21	
Total			22568	20572	17709	22568	21944	15556	20572	21944	15301

[a] Numbers as in 3.2.5

[b] Excluding those genes that were classified as IOs but are located in different synteny blocks

[c] Groups of genes in which several genes hit each other as RTHs

Table 3.3: Gene classification for each of the species pair comparisons, as well as the number of gene groups in each category excluding all genes from groups where one of the group members is outside of the synteny blocks.

			H. sapiens			***H. sapiens***			***P. troglodytes***		
			P. troglodytes			***M. mulatta***			***M. mulatta***		
Gene type[a]			**Hs**	**Pt**	**Gr**	**Hs**	**Mm**	**Gr**	**Pt**	**Mm**	**Gr**
Orthologues	1	Initial	15915			13310			13203		
		Synteny	364			393			312		
	4	Transposed[b]	23			110			125		
		Total	16302			13813			13640		
Duplications	2	Tandem	405	328	178	435	375	223	317	308	172
	3	Dispersed	2131	2005	843	1988	2147	893	1840	2089	848
Single genes	5		644	136		1748	347		1358	487	
Uncertain	6		1098	471		1645	2738		1367	2982	
Excluded	7		1988	1330	323	2939	2524	549	2050	2438	575
Total			22568	20572	17709	22568	21944	15556	20572	21944	15301

[a] Numbers as in 3.2.5

[b] Excluding those genes that were classified as IOs but are located in different synteny blocks

Table 3.4: Number of orthologue pairs shared between each pair of compared species, as well as 1:1:1 orthologue trios.

Species pair	Orthologue pairs[a]
H. sapiens & *P. troglodytes*	16376
H. sapiens & *M. mulatta*	13975
P. troglodytes & *M. mulatta*	13850
H. sapiens, *P. troglodytes* & *M. mulatta*	13274

[a] This number is the sum of the IOs and the synteny orthologues. Some of the transpositions were not in the original IOs and because they are translocated we have no synteny information on them. The IO where both members of the pair were out of the synteny blocks were excluded from table 3.2, and those transpositions which were not IOs are excluded from this table. This is the reason why the number of shared orthologues between each pair of species is slightly different than the sum of all the orthologues in table 3.2.

The enrichment of these GO categories in our sets of orthologues agrees with their observation. The GO category *transcription factor activity* includes both developmental transcription factors which are are highly conserved in mammals as well as the less well conserved non-developmental transcription factors, which explains the over-representation of this category.

When examining the distribution of the orthologue pairs among the different chromosomes, we found fewer than expected were located on the X and Y chromosomes, in both human and chimpanzee. This was also the case for the macaque X chromosome. The Y chromosome has not yet been sequenced in the macaque, so we have no information on it. This reduced number of orthologues observed is consistent with a higher mutation rate in the Y chromosome (CSAC, 2005) and in the case of the X chromosome with the three times higher rearrangement rate that has been observed in this chromosome (Gibbs *et al.*, 2007), as this could make it more difficult to use synteny information for the identification of orthologues that are not clear, which could result in fewer being identified in this chromosome.

3.4.2 Duplications

Those cases in which a group of more than two genes had Reciprocal Top Hits (methods on page 60) at the expected location and no hits out of that area were classified as tandem duplications. The genes in these groups may be duplicated in only one of the species (one-to-many relation) or in both (many-to-many relation). Those cases in which a group of more than two genes had RTHs at the expected location and at least one of them had a hit outside of the expected location were considered dispersed duplications. The number of groups of each kind we found can be seen in table 3.2, as well as the number of genes from each species within these groups. An example of tandem duplication can be seen in the MLRM and MRCL2 in figure 3.8 on page 81, in this figure there are also some examples of dispersed duplicates.

A very large number of duplicated gene groups are obtained from the comparison between human and macaque when compared to the number obtained from the chimpanzee vs. macaque comparison. This could be explained by recent duplications missing or being collapsed in the chimpanzee and macaque assemblies (She *et al.*, 2004), which would artificially inflate the number of duplications observed in human when compared to the other two primate species. This effect may also be responsible for the similar number of duplications observed in the human vs. chimpanzee and chimpanzee vs. macaque comparisons, when

Table 3.5: Differences in number of members of groups of highly similar genes.

Genome A	Genome B	Group class	Equal size	Size diff $= 1$		Size diff ≥ 2	
				$A > B$	$A < B$	$A > B$	$A < B$
H. sapiens	*P. troglodytes*	Tandem	107	60	3	8	0
		Dispersed	784	150	28	61	4
H. sapiens	*M. mulatta*	Tandem	61	96	56	9	1
		Dispersed	659	164	221	47	82
P. troglodytes	*M. mulatta*	Tandem	55	55	54	6	2
		Dispersed	625	118	249	21	89

we would expect a much lower number in the first case due to the more recent divergence of the two lineages.

Genes classified as dispersed duplications are groups of close paralogues that have been caused by duplications and translocations. A manual examination showed that many could be divided into pairs of one-to-one orthologues. They were not resolved into orthologue pairs by our automated method as we require both similarity and synteny information in order to prevent incorrect orthology assignments when the relation is not clearly one-to-one. Because of the high similarity between these sequences any attempt to resolve them would rely entirely on synteny information, which without manual curation could produce erroneous assignments. These groups would include both cases in which a gene has been duplicated and the duplicated copy relocated in another area of the genome prior to the divergence of the two compared species, and cases in which one of the genes has been duplicated after the divergence. However, we cannot know from this classification if a gene belonging to one of these groups has translocated in one of the species since the time of divergence, or before.

The GO analysis of the tandem duplicated genes shows *electron transport* and *oxygen binding* as significantly over-represented in all cases. In the case of those dispersed duplicates there were many more terms with a distribution that was significantly different from expected. Terms that were significantly over-represented included those related to *translation*, *ion transport* and *cytoskeleton activity*. These remained significant when the analysis was repeated selecting groups of genes instead of individual genes to avoid bias caused by all genes within a group having similar GO annotations.

To determine if there was any pattern in those groups of genes with a different number of genes in each of the species we separated these groups into five categories: those with the same number of genes in both species, those with a difference of one gene (two categories), and those with more than one gene difference (two categories). The number of genes in each category can be seen in table 3.5 on page 78. When examining the GO term distribution of the human genes from the groups in each of these categories, in the tandem duplicated groups, *oxygen binding* was over-represented in those groups with more members in human. In the dispersed duplicates groups, the terms *chromosome* and *translation* were significantly over-represented, with more human genes. *Translation* was also over-represented in groups with fewer genes in human.

The number of genes belonging to the human X chromosome in the tandem duplication category is significantly higher than expected in the comparisons of human with both chimpanzee and macaque. This observation may be explained by the presence of several clusters of five or more genes involved in recent gene duplication on this chromosome (IHGSC, 2004).

3.4.3 Single gene relocations

We identified 256 groups of orthologues that were in different synteny blocks. This number is influenced by the fragmentation of the assembly in the two compared genomes. A greater fragmentation will cause an inflation of the number of observed translocations. This will happen if an area where the synteny is conserved is broken by an assembly gap, as two genes that in reality are within each others expected location could end up in different synteny blocks and because of this be classified as a translocation. Although this is unlikely to occur frequently we have to interpret the results with care. Those genes in which one or both of the members of the pair are outside of the synteny blocks were also excluded because we cannot be sure a translocation and not an assembly problem is responsible for their placement outside of the blocks. An example of translocation can be seen in figure 3.8 overleaf for gene DCP2 (mRNA decapping enzyme) which has been translocated from chromosome 5 to chromosome 18 in the chimpanzee lineage.

In order to estimate the frequency of translocation in each of the lineages we need to determine in what lineage each of these translocations occurred. Because of this only those cases in which there was a clear relation between three genes,

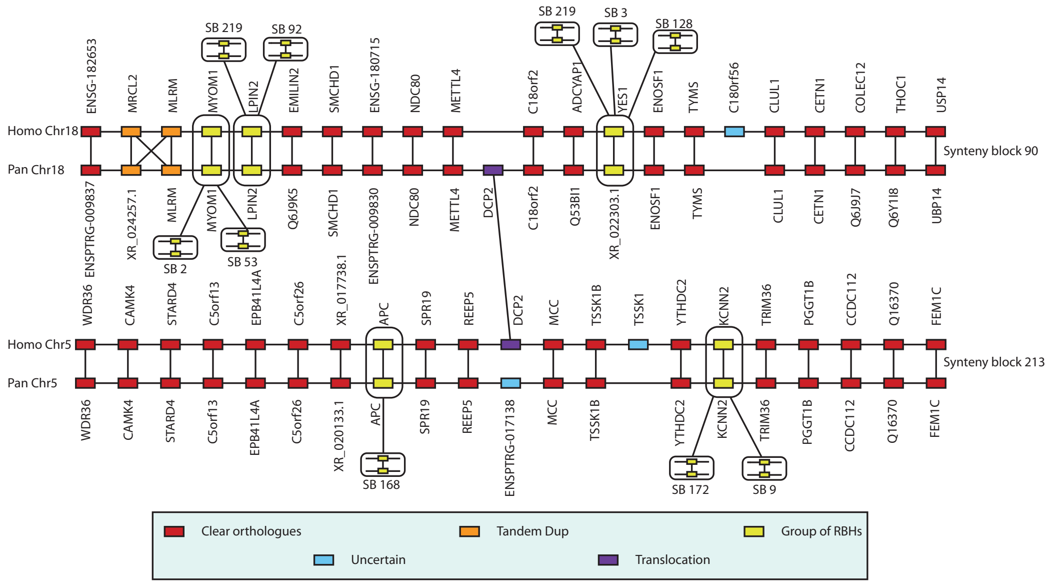


Figure 3.8: This figure shows an example of a translocated gene as well as the genes surrounding the affected gene DCP2. If no other name was assigned to the gene the EnsEMBL ID was used, but for reasons of space it was shortened with a "–" representing 00000. Descriptions available for the genes in the figure are shown in table 3.6 overleaf.

Table 3.6: This table shows descriptions available for the orthologues and groups of RTHs shown in figure 3.8 on the page before.

	Synteny block 90		Synteny block 213
Gene Hs/Pt	Description Hs / P t	Gene Hs/Pt	Description Hs / Pt
ENSG-182653 / ENSPTRG-009837	none	WDR36	WD repeat protein 36
MRCL2 / XR_024257.1	myosin regulatory light chain	CAMK4	Calcium/calmodulin-dependent protein kinase type IV
MLRM	Myosin regulatory light chain 2, nonsarcomeric	STARD4	StAR-related lipid transfer protein 4
MYOM1	Myomesin-1	C5orf13	Neuronal protein 3.1
LPIN2	Lipin-2	EPB41L4A	Band 4.1-like protein 4A
EMILIN2 / Q6J9K5	Elastin microfibril interface-located protein 2	C5orf26	Uncharacterized protein TIGA1
SMCHD1	Structural maintenance of chromosomes flexible hinge domain-containing protein 1	XR_017738.1 / XR_020133.1	none
ENSG-180715 / ENSPTRG-009830	none	APC	Adenomatous polyposis coli protein
NDC80	Kinetochore protein Hec1	SRP19	Signal recognition particle
METTL4	Methyltransferase-like protein 4	REEP5	Receptor expression-enhancing protein 5
- / DCP2	mRNA-decapping enzyme 2	DCP2	mRNA-decapping enzyme 2
-	-	ENSPTRG-017138	none
ENSG-132204 / ENSPTRG-009826	none	MCC	Colorectal mutant cancer protein
ADCYAP1 / Q53BI1	Pituitary adenylate cyclase-activating polypeptide	TSSK1B	testis-specific serine kinase 1B
YES1 / XR_022303.1	Proto-oncogene tyrosine-protein kinase Yes	TSSK1 / -	Testis-specific serine / Threonine-protein kinase 1
ENOSF1	rTS β protein	YTHDC2	YTH domain-containing protein 2
TYMS	Thymidylate synthase	KCNN2	Small conductance calcium-activated potassium channel protein 2
ENSG-176912 / -	none	TRIM36	Tripartite motif-containing protein 36
CLUL1	Clusterin-like protein 1 precursor	PGGT1β	Geranylgeranyl transferase type-1 subunit β
CETN1	Centrin-1	CCDC112	Coiled-coil domain containing 112 isoform 2
COLEC12 / Q6J9J7	Collectin sub-family / Collectin placenta 1	CTNNA1	none
THOC1 / Q6Y1I8	THO complex subunit 1	FEM1C	Feminization 1 homolog a
UBP14	Ubiquitin carboxyl-terminal hydrolase 14	-	-

one from each species, were considered. Within the list of 256 groups we identified 87 as 1:1:1 orthologues. One of these did not have conserved synteny in any of the genomes so the lineage in which it had occurred could not be determined. For the other 86 groups we inferred the branch in which the translocation had occurred by comparison of the gene order in the three genomes (figure 3.9 overleaf). The number of translocations on each branch is roughly proportional to the time.

There is still a large fraction of translocations that we cannot assign to any lineage, some of them because of missing data in one of the genomes, and others because the history of the gene appears more complicated that a simple translocation, though the ratios in each of the lineages is unlikely to be biased

We examined the distribution of the GO terms among the genes classified as translocations in each pair comparison (table 3.2 on page 75) however, we found no term with a significantly different frequency from expected in both of the comparisons.

In order to identify any areas of the genome where genes may translocate more frequently than expected, we examined the distribution of all the cases classified as translocations along the chromosomes and found no significant difference from expected except for the case of macaque chromosome 1 where there were significantly fewer translocations than expected when compared to both human and chimpanzee. This lower number of translocations in chromosome 1 was observed also in the human vs. chimpanzee comparison, however in this case the reduction was not statistically significant.

3.5 Candidate extinctions and creations

Those genes with no BLASTp hit are the best candidates for extinction/creation, however in order to make sure the absence of a BLASTp hit could not be explained by some other reason we performed a series of checks on these extinction/creation candidates.

We define gene creation as the appearance *de novo* of a gene in a species,

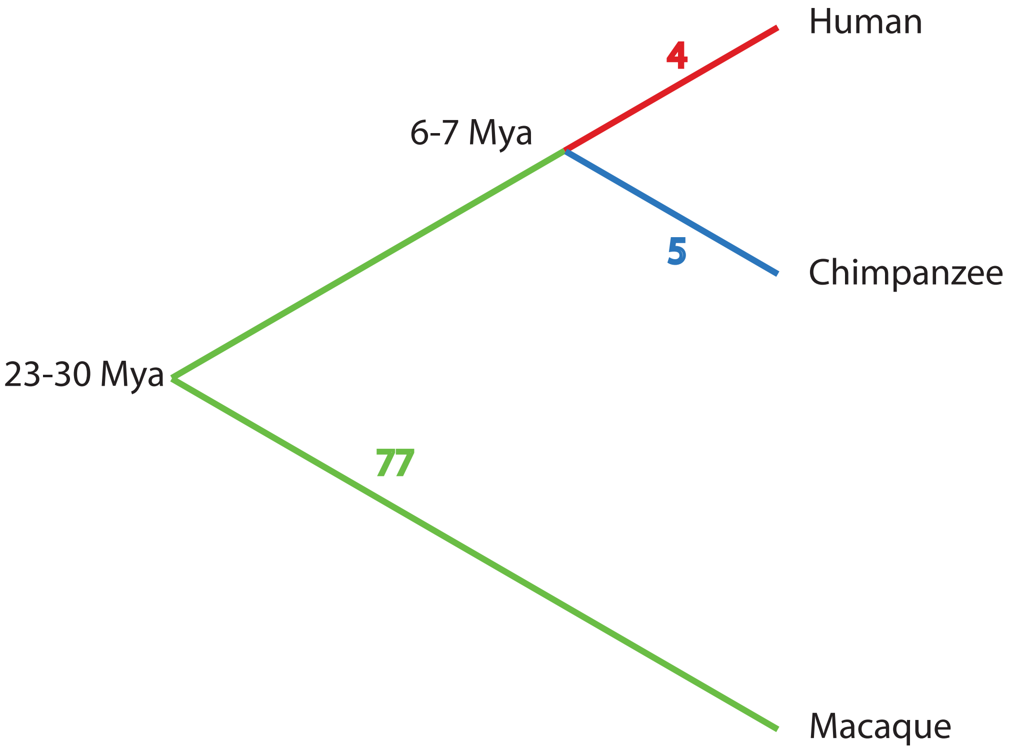


Figure 3.9: This figure shows those lineage specific translocation we identified that have occurred since the divergence of the three primate lineages. The numbers are coloured according to the branch on which the translocations occurred

where that gene has no similarity to any other gene in the genome. Gene extinction is defined as the loss of the last gene of a family (Demuth *et al.*, 2006). Genes that have no BLASTp hits in the other species are good candidates for gene extinction/creation events, in species with such short divergence times as human and chimpanzee. If the gene had originated by gene duplication since the speciation event, the absence of a BLASTp hit in the other species would mean all copies had been lost also since the speciation, or the gene had evolved so rapidly it is no longer recognizable, and both of these events seem unlikely. However the absence of a BLASTp hit on is own is not enough to guarantee they are extinctions or creations. To decide if they are real extinction/creations we first examined possible reasons that could lead to the absence of a BLASTp hit, and what we would expect to find for each of these cases:

- The gene is an annotation artifact. In this case we may find sequence similarity in the other species, but we would not expect to find any EST evidence for the gene, or conservation of the ORF.

- The orthologue is present in the other genome at the expected location, but the BLASTp search finds no similarity due to masking of the protein. In this case a comparison with the proteins encoded by all genes within the expected location should reveal a highly similar protein.

- The orthologue is present in the other organism, but missing in the genome annotation or in the assembly. In this case we should find a similar sequence at the expected location without frameshifts or stop codons within its ORF or an assembly gap corresponding to the area syntenic to the gene.

- The orthologue is a true loss in the other organism. We may see some remaining similarity if the loss occurred by inactivation and accumulation of deleterious mutations, or partial deletion in the other genome. If, however, it was by complete deletion of the area containing it, we would find no similarity.

- The gene is indeed a true gain. If it is an exaptation of existing intergenic sequence we expect to find certain nucleotide similarity with a close species in the syntenic area, but the ORF may contain frameshifts and/or stop codons. We investigate this possibility in chapter 4 on page 107.

In order to decide which of these cases is the most likely for each extinction/creation candidate, we used a series of different tests aimed to address each one of these issues (table 3.7 overleaf).

All those cases where there was a gap the size of the gene or larger at the expected location area were removed, as well as those where there was an inversion at the expected location area, as the absence of the gene would be uncertain. We used ssearch with a threshold of $1e \times 10^{-4}$ and BLAT in order to search for sequence similarity at the expected location, the results can be seen in figure 3.10 overleaf. The most promising extinction/creation candidates from this initial examination would be those with no gap and no hit at the expected location in the other species. However, the presence of a hit at the expected location does not necessarily indicate that the gene is there, it could also indicate the presence of a pseudogene or an exaptation event. For example the human gene Q4G0G9 which is annotated as an EnsEMBL KNOWN gene is located in an area where the gene order is perfectly conserved in chimpanzee and there are no large sequencing gaps or large sequence similarities at the expected location of the chimpanzee orthologue (figure 3.11 on page 89).

Even from this preliminary examination we would like to point out the big difference between the fraction of human genes with no BLASTp hit that have a gap at their expected location in the comparisons with both chimpanzee and macaque (66% and 77% respectively), and the fraction of both chimp and macaque genes with gaps at their expected location in the human genome (6% and 5%) respectively. When comparing macaque with chimpanzee, the fraction of genes with a gap at the expected location in the other species is larger for both organisms (76% of the chimpanzee and 59% of the macaque genes) and more similar when

Table 3.7: Checks that were done on the extinction/creation candidates, and what each test is intended to detect.

Test	Aim
BLAST against ESTs database	Decide if the gene has supporting evidence
Presence of VEGA Havana annotation	
Annotated as EnsEMBL KNOWN	
Search for sequencing gaps	Assembly problems
Translated BLAT search against the nucleotide sequence at the expected location	Detect annotation problems
Smith-Waterman search against the nucleotide sequence at the expected location using ssearch	
High identity to a protein annotated at the expected location (but no BLASTP hit due to low complexity)	
Pseudogene hit in the other species	Signs of the loss of this gene in the other species

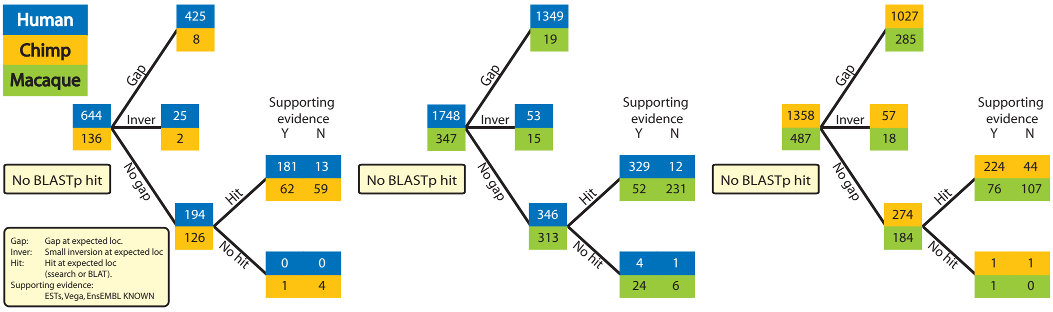


Figure 3.10: Classification of the genes according to the presence or absence of assembly gaps at the gene's expected location. For the cases with no gaps, these genes are further separated into those with and without ssearch or BLAT hits, and the number of genes within each category that have supporting evidence is indicated.

compared to each other. When examining the macaque comparisons we find the fraction of genes from the other two species with a gap at their expected location in this organism is higher, which is what we would expect from the more fragmented genome assembly. This observation is a very clear consequence of the difference between the finished state of the human genome and the draft state of the other two, and shows the need for careful checks when searching for differences in gene content between two genomes if one or both are in an unfinished state.

After this preliminary examination, those cases with a gap at the expected location were removed and we examined the remaining ones by aligning each of the protein sequences of the candidates with all proteins encoded by genes at the expected location. We removed those with an identity of 60% or greater to a protein within the area. These were not detected in the original BLASTp search because of masking. In order to detect any remaining cases, and also to ensure this method is adequate for detecting similar genes at the expected location, a BLASTp against the same group of proteins was carried out without masking. Only one case was found with an E-value lower than $1e-4$ that had not been previously removed by the protein comparison, and when examined it was found to be caused by a short area of high similarity. However, the overall sequence had an identity of less than 15%, and was not discarded. The 60% identity was chosen as the cutoff threshold based on the distribution of the identity values obtained from comparing each protein with all the proteins annotated at the expected location in each of the three species comparisons. This histogram (top part of figure 3.12 on page 91), shows a sharp drop in the number of hits with identity around 60% followed by a steep increase around 80% identity reflecting the number of cases where there is a highly similar orthologue in the region. This drop around 60% identity is likely to mark the end of the region of nonspecific identity. The greater conservation of synteny between human and chimpanzee is also reflected here by the larger number of genes with highly conserved sequences

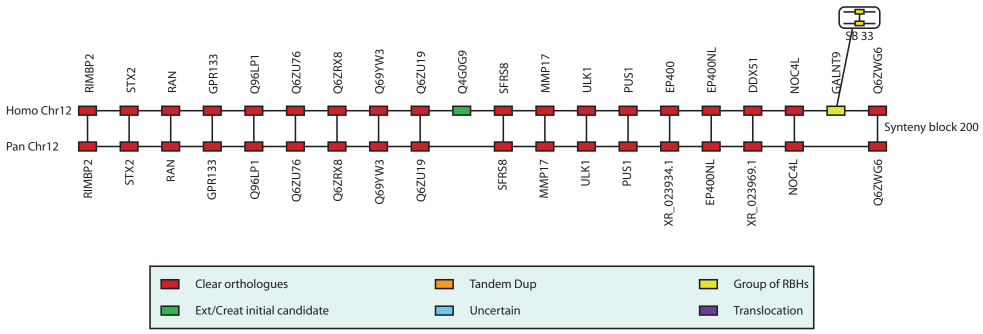


Figure 3.11: This figure shows an example of extinction/creation candidate as well as the genes surrounding it. Descriptions available for the genes in the figure are shown in table 3.8.

Table 3.8: This table shows descriptions available for the genes shown in figure 3.11.

Gene Hs	Gene Pt	Description[a]
RIMBP2	RIMBP2	RIM-binding protein 2
STX2	STX2	Syntaxin-2
RAN	RAN	GTP-binding nuclear protein Ran
GPR133	GPR133	Probable G-protein coupled receptor 133 precursor
Q96LP1	Q96LP1	none
Q6ZU76	Q6ZU76	none
Q6ZRX8	Q6ZRX8	none
Q69YW3	Q69YW3	none
Q6ZU19	Q6ZU19	none
Q4G0G9	-	none
SFRS8	SFRS8	Splicing factor, arginine/serine-rich 8
MMP17	MMP17	Matrix metalloproteinase-17 precursor
ULK1	ULK1	Serine/threonine-protein kinase ULK1
PUS1	PUS1	tRNA pseudouridine synthase A
EP400	XR_023934.1	Hs E1A-binding protein p400/Pt similar to KIAA1498 protein
EP400NL	EP400NL	EP400 N-terminal-like protein
DDX51	XR_023969.1	Hs ATP-dependent RNA helicase/Pt Hypothetical
NOC4L	NOC4L	Nucleolar complex protein 4 homolog
GALNT9	-	Polypeptide N-acetylgalactosaminyltransferase 9
Q6ZWG6	Q6ZWG6	none

[a] If the description for the gene is different in the two species the two annotations are separated by a ’/’.

at the expected location in the other species. When the same distribution is built for the initial candidates, in which case there is no orthologue annotated at the expected location, we still observe a similar shape in the histogram before the drop at around 60% identity, but no sharp increase in the higher identity region.There is a smaller increase caused by those low complexity proteins that were missed by the BLASTp and are indeed orthologues, but overall it is much less pronounced than the one observed for the whole genome data, as most of these genes have no similarity at the expected location (bottom part of figure 3.12 overleaf).

Of the remaining proteins one had an orthologue annotated by EnsEMBL and was also removed. After this we used BLAT in translated mode in order to detect any similarity of the candidate protein sequence within the expected location area. Those cases with the same number of coding exons, no stop codons in frame and an identity of 90% or more for each of the coding exons at the protein level were discarded as possible misannotations where the gene may indeed be present in the other organism. A summary of all the checks performed for the comparison between human and chimpanzee can be seen in table 3.9 on page 92, the same can be seen for the comparison of these two species with macaque in tables 3.10 on page 93 and 3.11 on page 94.

From theses tables we can immediately see the huge difference in the fraction of genes belonging to human that have support (93% of the initial number of candidates) when compared to the same fraction in chimpanzee (50%), this is another clear indication of the difference in quality of the information available for the two species. The same is true for the human-macaque comparison where 96% of the initial human genes have support but only 19% of the macaque genes do.

In some cases the hits that the remaining genes had at the expected location in the other species were annotated as pseudogenes, this was true for 93 of the human genes when compared to chimpanzee, while only 6 of the chimpanzee

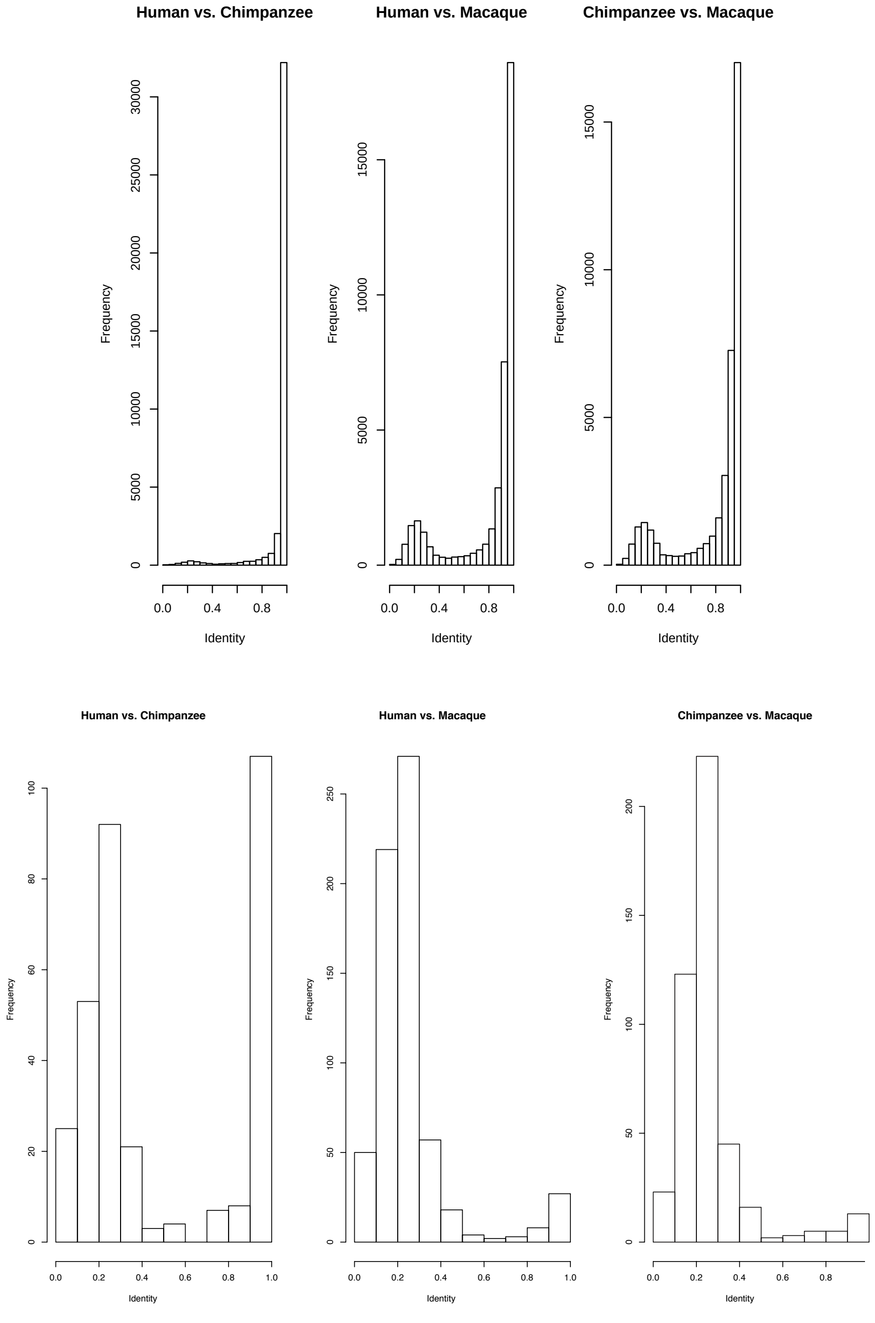


Figure 3.12: The top three histograms in the figure show the distribution of the identity values resulting from aligning each of the genes with no gap at the expected location with all of the proteins annotated at this location for each of the three pairs of species. The lower three histograms show the same distribution of identity values for the subset of initial candidates in each of the three pairs of compared species.

Table 3.9: Summary of the final filtering of the extinction/creation candidates obtained from the human-chimpanzee comparison.

Species	Hit at expected location[a]	Support[b]	Initial number	Genes remaining after the check	
				Similar protein at expected location	BLAT hit
H. sapiens[c]	Yes	No	13	3	3
	Yes	Yes	181	149	135
P. troglodytes	No	No	4	2	2
	No	Yes	1	1[d]	0
	Yes	No	59	32	15
	Yes	Yes	62	10	4

[a] BLAT hit, ssearch hit or both

[b] A gene is considered to have support if it is annotated as EnsEMBL KNOWN, if it has EST support or if it has Vega/Havana annotation

[c] All human genes had a hit at the expected location in chimpanzee

[d] This gene is the only one that had an orthologue annotated by EnsEMBL, so it was removed from the set

Table 3.10: Summary of the final filtering of the extinction/creation candidates obtained from the human-macaque comparison.

Species	Hit at expected location[a]	Support[b]	Initial number	Genes remaining after the check	
				Similar prot. at expected loc. or EnsEMBL orthologue[c]	BLAT hit
H. sapiens	No	No	1	1	1
	No	Yes	4	4	4
	Yes	No	12	12	12
	Yes	Yes	329	314 (2)	297
M. mulatta	No	No	24	17 (7)	17
	No	Yes	6	3 (3)	3
	Yes	No	231	214 (9)	161
	Yes	Yes	52	30 (3)	22

[a] BLAT hit, ssearch hit or both

[b] A gene is considered to have support if it is annotated as EnsEMBL KNOWN, if it has EST support or if it has Vega/Havana annotation

[c] the cases that were removed because of the presence of an annotated EnsEMBL orthologue are indicated in brackets

Table 3.11: Summary of the final filtering of the extinction/creation candidates obtained from the chimpanzee-macaque comparison.

Species	Hit at expected location[a]	Support[b]	Initial number	Genes remaining after the check	
				Similar prot. at expected loc. or EnsEMBL orthologue[c]	BLAT hit
P. troglodytes	No	No	1	1	1
	No	Yes	1	1	1
	Yes	No	48	44 (0)	39
	Yes	Yes	224	213 (2)	205
M. mulatta	No	No	0	0	0
	No	Yes	1	1	1
	Yes	No	107	100 (4)	76
	Yes	Yes	76	63 (3)	49

[a] BLAT hit, ssearch hit or both

[b] A gene is considered to have support if it is annotated as EnsEMBL KNOWN, if it has EST support or if it has Vega/Havana annotation

[c] the cases that were removed because of the presence of an annotated EnsEMBL orthologue are indicated in brackets

genes hit a human pseudogene (table 3.12).

Those genes that passed all checks in each of the species were compared with the macaque genome in order to determine if we could assign them as lineage specific gains or losses (table 3.13 overleaf).

We examined those 9 macaque and 32 human lineage specific extinction/creation candidates that had supporting evidence in search for any indication of their function (table 3.14 overleaf). Most of them had no annotation that would allow us to determine their function, which would be expected from new genes that may have arisen by exaptation of non-coding sequence, however it would also be expected from annotation errors introduced by the gene prediction methods. We examined the GO term distribution of all the extinction creation candidates obtained in the comparison of each species with the other two, but the results obtained are not conclusive, as most of them have no GO annotation.

In the case of the 33 human and the 1 chimpanzee lineage specific genes, they are not present in an out-group (macaque) either, so they are good candidates for gene creation events in the respective lineages. The possibility of parallel loss exists also in all cases, but we consider this quite unlikely. These creation candidates are examined in more detail in chapter 4.

In the case of the lineage specific genes in macaque we cannot determine if they are gene creations in macaque or if they are genes that were present in the common ancestor of Hominidae and Cercopithecoidea that have been lost

Table 3.12: Summary of the number of extinction/creation candidates, that show hits to a sequence annotated as a pseudogene in the other genome. The number shown in brackets is the total number of candidates that passed the filters.

Species	Species it is compared with		
	H. sapiens	*P. troglodytes*	*M. mulatta*
H. sapiens	-	93 (138)	26 (314)
P.troglodytes	6 (27)	-	21 (246)
M. mulatta	17(203)	10 (126)	-

Table 3.13: Summary of the number of lineage specific gene creation candidates. The two numbers shown are the total number of genes and the number of genes that have supporting evidence (Annotated as EnsEMBL KNOWN, have EST support or have VEGA/Havana annotation).

Species	Lineage specific[a]
H. sapiens	33/32
P.troglodytes	1/0
M. mulatta	65/9

[a] Number of genes classified as extinction/creation candidates in the comparison with both of the other species

Table 3.14: Summary of the functional information available for the lineage specific genes with supporting evidence.

Species	Genes	Annotaton	Goslim
M. mulatta	1	Elastin	None
	1	Uricase (EC 1.7.3.3) (Urate oxidase)	Peroxisome
	7	No description	None
H. sapiens	1	Nervous system abundant protein 11	None
	1	PRKR interacting protein 1 (IL11 inducible)	None
	1	Chronic lymphocytic leukaemia up-regulated 1	None
	1	Complexin 1 (CPLX1)	None
	1	Tumor suppressor candidate 5 (TUSC5)	None
	1	Putative proline-rich protein DAMS	Signal transduction
	26	Putative protein/No description	None

in the lineage leading to Hominidae before the divergence between human and chimpanzee. Because of this we performed no further analysis on this group of genes in this work. However, they merit further investigation when more primate genomic information is available for comparison. In any case, it is gratifying that this group of genes included the Urate oxidase, a gene that has been identified previously as an ape specific gene inactivation (Wu *et al.* 1989, Oda *et al.* 2002), as this indicates our pipeline is working correctly.

When examining the chromosome distribution of the extinction/creation candidates, there was a significantly greater number of these genes located in the human and chimp chromosomes 20 and 21 than expected when comparing to macaque. When comparing human vs. chimpanzee there is a highly significant excess of genes belonging to human chromosome 21 in the group of extinction creation candidates. This chromosome has been studied in detail (Reymond *et al.*, 2002), and this may have produced a higher quality annotation which could explain the excess of human annotated genes when compared to the other two primates.

3.6 Discussion

What makes us human is a question mankind has been trying to answer since the beginning of science. The completion of the human genome (Lander *et al.*, 2001) was a big step towards this end, but there is still a long way to go. The sequencing of the chimpanzee (CSAC, 2005), our closest living relative, and the macaque (Gibbs *et al.*, 2007), which is distant enough to make the identification of conserved regions of biological significance easier have been other great steps on the way. However, identification of those particular differences that cause our species phenotype is still to be achieved.

Assuming human and chimpanzee diverged 6 Mya and their lineage split from macaque around 30 Mya, if the genome data is complete we should be able to estimate the rates at which genes are created and become extinct in the primate lineage, as well as the frequency of translocation events, and thus shed a little more light on the differences between ourselves and some of our closest neighbours.

An overview of the three genomes shows, as has been noted before (Gibbs *et al.*, 2007), a great degree of conservation at the higher level, with most of the genome in all three species contained within synteny blocks. The simplest explanation for the greater number of synteny breaks between chimpanzee and macaque than between human and macaque is the lower degree of completion of these two genomes, as there is no indication in the literature of a huge difference in the frequency of genome rearrangements since the divergence of these two species that would otherwise explain this difference (Cheng *et al.* 2005, Kehrer-Sawatzki and Cooper 2007a).

The fraction of shared orthologue pairs found in the human vs. macaque and chimpanzee vs. macaque comparisons are similar, the percentage of macaque genes with orthologues in each comparison being 63.7% and 63.1% respectively. This observation agrees with what we would expect if human and chimpanzee have a similar rate of evolution. The fraction of shared orthologues obtained between human and chimpanzee is higher, with 79.6% chimpanzee genes and 72.6% human genes having orthologues in the other species. This is also what we would expect from their shorter divergence times.

When observing the fraction of genes in each genome classified as duplications (both tandem and dispersed), the fraction is similar in the three species, ranging from 13.7% of the genes in chimpanzee to 14.5% in human. The small difference between them could also reflect differences in the genome coverage, as human with the finished genome has the highest number, and chimpanzee with the lowest coverage has the smallest. This is also the case with the number of genes annotated for each of the genomes.

There are more tandem duplicated groups found between human and both chimpanzee and macaque than between macaque and chimpanzee. This observation is not what we would expect from the phylogeny of these species, but again it can be explained by the higher quality of the human genome.

One of the problems that have been noted in shotgun genome sequencing is the inability to resolve correctly segmental duplications that are $> 97\%$ identical (She *et al.*, 2004) which would include most recent duplications. It has been estimated that 5% of the human genome is formed by segmental duplications that originated in the past 35 My (Samonte and Eichler, 2002). If these recent duplications contain genes, and have been collapsed in the lower quality assemblies, we would expect an artificial increase in the fraction of observed tandem duplication where the number of genes in the duplicated group is larger in the genome with the better quality. This effect should not be so large when comparing chimpanzee and macaque, as segmental duplications are likely to have been collapsed by the same degree in both of the genomes. Indeed when we examine table 3.5 on page 78 we can see the number of tandem duplications that contain more genes in human when compared to both chimpanzee and macaque is much greater than the number that contain more genes in the opposite species, while when comparing the chimpanzee and macaque genomes the numbers are very similar. The incompleteness of the assembly offers us a far simpler explanation than alternative scenarios that would require a huge increase in the tandem duplications in the human lineage since its divergence from chimpanzee.

Genes that are classified as dispersed duplicates will be affected by two genome quality dependent factors, the collapse of segmental duplications noted above, and the fragmentation of the assembly. These two factors will act in the opposite way, the fragmentation of the assembly will inflate the number of groups with more genes in the species with a more fragmented assembly. This is because any real tandem duplication that involves a gene located in unfinished regions of the assembly (EnsEMBL chromosome random or unknown regions, and scaffolds) will cause an artificial synteny break. This effect depends on the number of genes from unfinished regions that are included in the groups. In the case of the human vs. chimpanzee comparison there are genes belonging to 21 chimpanzee unfinished regions, in the human vs. macaque comparison there are genes from 98 macaque unfinished regions, and in the chimpanzee vs. macaque there are genes from 96 unfinished regions (19 from chimpanzee and 77 from macaque). According to this we would expect a greater inflation of the number of macaque genes in groups classified as dispersed duplicates, followed by a lower inflation of the chimpanzee genes, which is what we observe (table 3.3 on page 76). We can also see in table 3.5 on page 78 that the number of groups classified as dispersed duplicates in which there are more macaque genes when compared to the other two species is greater than the number of groups in which there are more genes in the other species.

The number of translocations obtained from each species pair comparison is not affected by the assembly fragmentation, as we only included those pairs of translocated genes in which both members belong to one of the synteny blocks. However, this means that the number of translocations we obtain is just a lower boundary for the real number.

In order to determine if this rate has changed since the speciation of the Hominoidea, we need to estimate the rate of translocations before their divergence. The 77 translocations that are different in macaque from both human and chimpanzee are the total number of translocation events that have occurred over two evolutionary branches originating at the time when the Cercopithecoidea diverged. One of these branches lasted $\approx$ 30 Myr and led to the macaque lineage, the other $\approx$ 24 My and led to the point where human and chimpanzee diverged. The simplest estimation would be to assume a constant rate over time and calculate the number of translocations per My from there. However, the number of translocations per million years does not reflect the differences in the number of generations that may have occurred along these two branches. The number of observed differences are those that have occurred on the germ line and a longer generation time will mean fewer replications along the germ line during the same period of time. A slowdown in the molecular clock in humans and to a lesser extent in chimpanzee has been observed and related to this longer generation times in these species (Elango *et al.*, 2006).

We calculated the number of translocations per generation that would have occurred in human, 1.33×10^{-5}, and in chimpanzee 1.25×10^{-5} assuming a generation time of 20 years for human and 15 for chimpanzee (Elango *et al.*, 2006). In the case of the 77 translocations that are differences between macaque and the hominid lineage, 6×10^{6} macaque (generation time 5 years) generations would have occurred in 30 My and 1.6×10^{6} hominid generations. We use the chimpanzee generation time as an estimate of the generation time along this branch, as the increase in the generation time in humans occurred in the last 6 My, so the chimpanzee will more accurately reflect the ancestral generation time. These generation times are used as there appears to have been an increase in the generation time in the Hominidae lineage with both gorilla and orangutan having longer generation times than primates from earlier diverging lineages such as lemur, macaque and baboon which all have have a generation time close to 5 years. We are assuming the difference in generation time occurred shortly after the speciation event and the translocation rate per generation is the same along both branches. If these assumptions are correct, the rate of translocation per generation along both of these branches is 1.01×10^{-5}. These estimates all depend on the generation times of the ancestors of the examined species which are uncertain, as well as the divergence times, however they show no difference in magnitude that may have indicated a large difference in the rates between these different species.

Regarding the extinction/creation candidate search, we will examine it in more detail in chapter 4 on page 107, however we will summarize briefly some of the main results we obtained. We noted a great number of human genes that have a gap at the expected location in the chimpanzee as well as a greater fraction of chimpanzee genes with no supporting evidence, this reflects the WD status of the chimpanzee genome and also the huge difference in the effort invested in study of these two species. This is clearly reflected in the size of the EST data available for each of them, 6930 sequences available in chimpanzee versus nearly 8 million in human. The same results are found in the comparison of human with macaque, although the EST coverage for this organism is better with 60 thousand ESTs from *M. mulatta*, and a further 140 thousand from other *Macaca* species.

We identify 9 genes that may have been lost either in the Hominidae branch, before the divergence of the lineages leading to human and chimpanzee, or in the branch leading to the macaque. Of these it is worth noting one of them, the Urate oxidase is a well known case of ape specific gene loss (Oda *et al.*, 2002). This confirms our approach works correctly, at least in the measure that it is able to identify a known lineage specific gene loss, although it cannot assign it to a specific branch because of the lack of an outgroup.

Although we can identify 33 candidates for exaptation in the human genome and only one in the chimpanzee, this does not mean this process is occurring more frequently in humans. We eliminated all those candidates that were not annotated as Known by EnsEMBL or had EST support in order to remove false positives. This means we have removed those chimpanzee and human cases that were only supported by gene prediction algorithms, and some of these cases could be real genes. This would be particularly relevant in the case of the chimpanzee genes, as there is much less information available for this species than for human that would validate gene predictions. This means that in this step more chimpanzee than human genes were removed. Another issue is that because the chimpanzee genome relied heavily on the human genome for its assembly and annotation it is more likely that a gene with no orthologue in human would remain un-annotated in chimpanzee than the opposite.

The large number of genes from human with gaps at their expected location when compared to chimpanzee and macaque contrasts sharply with the small number of genes from each of these organisms with a gap at their expected location in human, and also with the numbers obtained for the chimpanzee versus macaque comparison where the fraction of genes belonging to each organism with a gap in the other one is much more similar.

This difference in the quality of the genomes is an important factor affecting the results of any study aimed to find a complete set of differences between two genomes, and it will affect the obtained results in most cases by inflating the apparent differences between the genomes. However by using a synteny framework we can identify many of these cases where assembly or annotation problems may produce artifactual gene content differences.

To obtain a complete overview of the differences in gene content between human and chimpanzee, we considered both those differences caused by changes in gene copy number between the two species and those cause by the extinction/creation of genes in one of the lineages. We can obtain an estimate of the differences caused by changes in copy number by using the groups of genes classified as duplications. We found 891 groups (excluding 1:1 orthologues), comprising 4486 genes where the number of members in the group is equal between human and chimp (table 3.5 on page 78). In 210 groups and 31 groups there is one more gene in human or chimpanzee respectively – corresponding to only 241 gene indels. Furthermore, 69 groups and 4 groups have at least two genes more in human and chimpanzee respectively, amounting to at least 220 gene indels since their common ancestor. This amounts to 461 gene indels caused by different copy number within gene groups between these two species.

There are an additional 159 potential differences between human and chimpanzee caused by gene extinction/creations (table 3.9 on page 92). This produces a total difference in gene content of around 2.8% (620 / 22500 genes) between human and chimpanzee, which is higher than the 1.5% difference in nucleotide sequence between orthologous regions (CSAC, 2005).

When comparing human with macaque and chimpanzee with macaque in a similar manner we found a difference of 6.7% and 5.3% respectively. The larger difference in the human-macaque comparison is probably another reflection of the finished state of the human genome.

The results of this study suggests a lower level of gene content divergence than suggested in previous studies. Demuth *et al.* (2006) suggested that 1418 human had no orthologue in the chimpanzee lineage while only 161 chimpanzee genes had no orthologues in the human, resulting in a difference of more than 6% in the gene content between the species. The number we obtain is much smaller. By using the synteny framework many of the human genes that had no similarity in chimpanzee were found to contain gaps at the expected location. The Demuth *et al.* (2006) data show a large bias towards gene family expansion in human and contraction in chimpanzee. A similar observation is also obtained in by Blomme *et al.* (2006) where the organisms that showed a greater number of gains and fewer losses were human and rodents. These are both predictable artifacts of the different levels of sequencing and annotation in the two genomes as we have seen in our analysis.

The results of the GO analysis of the different groups in most cases shows no clear biological interpretation for the differences in the GO term distribution observed in the different groups. In some cases, like *cell cycle*, *transcription* and *nucleus* that were over-represented in the orthologues, this could reflect a higher conservation of genes involved in fundamental processes as was noted by Lopez-Bigas *et al.* (2008), as well as the location of many of the proteins involved in these fundamental processes. *Protein transport* was also over-represented, which could indicate the conservation of those mechanisms involved in nuclear import and export. In the duplicated genes, *ion transport* was over-represented, which may reflect a greater plasticity of transporter proteins, and *translation* was over-represented also.

The differences in the location of genes belonging to the different groups correspond to what has been previously reported in the literature. Some chromosomes have suffered more rearrangements during their history, which reduces the synteny conservation between them. High proportions of segmental duplications in a chromosome are reflected in the number of tandemly duplicated genes that are located in it. However, before a comprehensive analysis of these rearrangements in human is possible, a finished quality genome with which we can compare it is required.

In our analysis we identified 13274 1:1:1 groups of orthologues that showed conserved synteny in the three species, this represents 59% of the human annotated genes. However this number may be increased even further by obtaining a better resolution of the dispersed duplicates, which are in many cases highly similar paralogues. In order to enable us to resolve these groups classified as duplications, phylogenetic analysis may be added to the analysis pipeline, which would give us a finer view of the differences in duplication rates between the species, although, we will still be limited by the genome quality.

From this study, we can extract several conclusions.

- The current quality of the available primate genomes, although very good for local comparisons and determining the general similarity between these species, is insufficient for an accurate determination of gene content difference between the complete genomes, or to determine if there are any significant differences in the pattern of gene duplications and translocations between the three species.

- The difference in quality alone between the finished genomes is sufficient to explain the greater number of gene gains and smaller number of losses that have been reported for the human lineage when compared to all other sequenced mammals.

- By using the data available and very strict criteria to define gene content differences between the three primate lineages we can only find a 2.8% difference between human and chimpanzee. However, this is a tentative value, that may change with the improvement of the genome sequences and annotation of the chimpanzee genome, as only a finished genome for both species would allow a complete determination of the differences between them.

Chapter 4

Exaptation of non-coding sequences to form new genes in the human genome

4.1 Introduction

The difference in gene content between human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*) can be caused by variation in the number of copies of the same genes, or by the complete absence of a gene family in one of the species. The presence of a gene in the human or chimpanzee lineages and its absence in the other species will indicate a creation or an extinction event in one of the lineages. We can distinguish between the two kinds of event and also determine in which lineage the event took place by using the presence or absence of the gene in the macaque (*Macaca mulatta*) genome. A creation event will be caused by the generation of a new gene by some process other than gene duplication, as gene duplication would generate a new copy of the gene, that would belong to the same gene family. An extinction would be caused by the loss of the last gene within a gene family (Demuth *et al.*, 2006).

The generation of new genes is generally attributed to the duplication of existing gene sequences, usually by segmental duplication of regions containing complete or partial genes. These duplicated regions may remain in tandem or be translocated to different parts of the genome. After the duplication the two copies of the gene may generate new genes by subfunctionalization or neofunctionalization, or one of the copies may be inactivated by deleterious mutations (Force *et al.* 1999, Lynch and Conery 2000). A new gene may also be generated by the complete or partial duplication of a gene and its translocation within the coding sequence of another producing a chimeric gene (Zhang *et al.*, 2006). Although the process of gene duplication has been the predominant source of new genes during evolution (Ohno, 1970), this does not mean that it is the only source (Long *et al.*, 2003).

Other sources that have been suggested are retroposition (Burki and Kaessmann, 2004), transposons (Kapitonov and Jurka, 2005), horizontal transfer (Bernstein *et al.*, 1996), *de novo* polymerization of nucleotides (as occurs in the immune system), or exaptation of a previously unused open reading frame (ORF) within existing coding sequence (overprinting) or from non-coding DNA (Long *et al.*, 2003). We will use the term exaptation for those new genes that have originated from the expression in one species of an ORF contained in a region of the genome that is non-coding in a sister species as well as in their last common ancestor.

New coding regions generated by overprinting are found in many overlapping genes and alternatively spliced variants, where the nucleotides providing the new reading frame already encoded another gene, and there are many examples of these cases in viruses (Keese and Gibbs, 1992) and prokaryotes (Delaye *et al.*, 2008).

Overlapping genes are common, and were first discovered in viruses where they are likely to be favoured by the constraints imposed by the need to retain a small genome (Keese and Gibbs, 1992), however they have also been described in prokaryotes (Delaye *et al.*, 2008) and in eukaryotes (Bernstein *et al.* 1996, Brosius 1999, Burki and Kaessmann 2004, Makalowska *et al.* 2007).

Overlapping genes reflect the creation of new genes from existing ORFs within the genome that were not previously used. All nucleotide sequences have redundant ORFs that are potentially coding, and form a pool of protein domains that could, given the right circumstances, become expressed (Keese and Gibbs, 1992). If a gene is already being expressed, these potential coding regions can become expressed easily by the introduction of point mutations that alter the frame, remove or create an alternative splice site or extend the un-translated region (UTR), indeed the same mechanisms responsible for the creation of new exons or genes can produce overlapping genes (Makalowska *et al.*, 2007). If the newly generated ORF is beneficial it may be retained, and if not it is likely to be lost, as is the case with any other new gene.

Overprinting specifically refers to the use of ORFs already present close to or within an existing expressed gene, but there are many other ORFs distributed along the genome, that could potentially become coding. In these cases there are additional difficulties, one of them being that the sequences lack the necessary signals for their transcription.

However, it is conceivable that in some cases a completely new ORF may arise from this pool of unused ORFs distributed along the genome. One mechanism by which this may happen is the acquisition of an ORF by a genomic region that is expressed but does not encode a protein, as it has been noted that a much larger fraction of the genome is transcribed in mammals than previously thought (Pheasant and Mattick, 2007). Alternatively the promoter region of a gene could be duplicated and translocated to another region of the genome close to an existing ORF, or an ORF that exists close to another gene may be able to use the promoter of that gene.

If the exaptation of an ORF has occurred since the divergence between the human and the chimpanzee, we expect to find a novel expressed gene with only one exon, and a very similar region in the other species, which would look like a pseudogene, although it would not be a true pseudogene, because it would never have existed as a gene in that species. If we also find a similar region in the macaque genome, which we are using as an outgroup in this analysis, that also looks like a pseudogene, the most parsimonious explanation for this observation would be the creation of a new gene, rather than the parallel inactivation of the gene in the other two species.

In this chapter we examine the human and chimpanzee genomes for cases in which this may have occurred.

4.2 Materials and Methods

4.2.1 Data

The genes examined were the 33 human and 1 chimpanzee genes obtained as potential creation candidates (chapter 3).

4.2.2 EnsEMBL Orthologue

The presence of an initial methionine and a plausible intron-exon structure was evaluated manually by using the online access to the EnsEMBL version 46 which was used in the analysis (`http://aug2007.archive.ensembl.org/`).

4.2.3 Exaptation alignments

The alignments between the exaptation candidates and the genomic sequence present at the expected location in chimpanzee and macaque was based on the alignment returned by MultiPipMaker (Schwartz *et al.*, 2000) and curated manually.

4.3 Results

We compared the complete set of protein coding genes from human (*Homo sapiens*), chimpanzee (*Pan troglodytes*) and macaque (*Macaca mulatta*) in search of cases in which we could classify a gene as specific to one of the two hominid lineages with a high degree of certainty (see chapter 3 on page 55). 34 genes were obtained that appeared to be good candidates for gene creations. These were genes that had no BLASTp hit in the other two species, no large gap (the size of the gene or larger) at the expected location, no similar protein annotated at the expected location, and no BLAT hit at the expected location in which each of the exons of the creation candidate were conserved with an identity $\geq 90\%$ without any in-frame stop codons (see section 3.5 on page 83).

There were also 65 cases that were specific to the macaque lineage. These could be caused by the creation of the gene in macaque or by a loss of the gene in the lineage leading to Hominidae, but we cannot determine which is the case from the comparison of these three primate genomes. Because of this we did not examine them further.

In the case of those genes that were specific to either human or chimpanzee, their presence could be explained either by two parallel losses, one in macaque and one in the other hominid, or by one creation event. A single creation event is the most parsimonious explanation however in order to verify this is the case we must be certain these genes are not present in the other hominid.

Classification of a gene as lineage specific could be caused by annotation errors or missing sequence in the genome where the gene appears to be absent. In order to rule this out we also checked for any evidence that the gene may be present in the other species but appear not to be there. We checked rigorously for any possibility of annotation artifact and small sequencing gaps that may affect the coding regions. We also checked for evidence of expression, to support the veracity of the annotated gene, as these sequences might have been predicted as coding because they contain a long ORF or biased base composition even if they are not expressed.

4.3.1 Initial Evaluation of the Candidates

Search for Assembly Gaps at the Expected Location

For each of the candidates the area containing the gene was aligned with the orthologous area in the other two species using MultiPipMaker (Schwartz *et al.*, 2000). This will reveal any gaps that may be present within the area containing the gene but would be smaller than the size of the gene, and so would not have been detected in the initial filtering.

Six cases were removed due to the presence of assembly gaps within the regions that aligned with coding exons in human. An example of this can be seen in figure 4.1 for the human PRKR interacting protein 1 (PRKRIP1) gene. This gene interacts with the interferon-induced dsRNA-activated protein kinase (PRKR) which belongs to a subclass of serine/threonine kinases, involved in the regulation of protein synthesis. This protein also has orthologues annotated by EnsEMBL in several other organisms as well as experimentally verified orthologues in mouse (*Mus musculus*) and rat (*Rattus norvegicus*).

Another seven cases in which there was a gap in the chimpanzee or the macaque sequence affecting only the non-coding area of one the predicted exons were kept. These were the genes ENSG00000180358, ENSG00000183633, ENSG00000204380, ENSG00000204601, ENSG00000206062 which had no annotation and ENSG00000184811 which is described as tumor suppressor candidate 5 (TUSC5).

Search for Deletions within the Coding Regions

We searched for deletions in the coding region of the candidates that may explain the absence of the gene in one of the species is due to the deletion of the affected gene even if there are no sequencing gaps. These deletions can be easily identified from the alignments produced by MultiPipMaker.

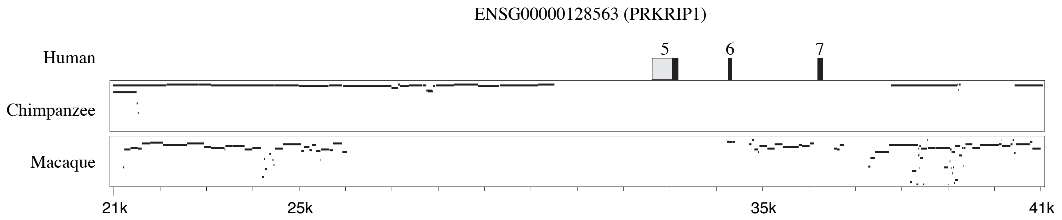


Figure 4.1: This figure shows a detail of the MultiPipMaker alignment between the human gene PRKRIP1 and the orthologous region in chimpanzee and macaque. The upper part shows the location of exons 5, 6 and 7 of human PRKRIP1 gene. The protein coding area is coloured in black and the UTR in grey. The lines within the two boxes marked as chimpanzee and macaque indicate the percent identity of the respective sequence when compared to human. This identity ranges from 50% at the bottom of the box to 100% at the top. There are sequence gaps in the areas where the black identity lines are missing, these gaps span exons 5, 6 and 7 in the case of chimpanzee and 5 and 7 in the case of macaque.

In the case of ENSG00000177627 we found that the gene sequence is missing in chimpanzee, although there is no sequencing gap in the syntenic region in the chimpanzee genome. There is a well conserved orthologue annotated in the bushbaby (*Otolemur garnettii*) and also in other more distant species. We also observed a good alignment at the nucleotide level between human and macaque (figure 4.2), this indicates the gene is probably present in macaque although not annotated. This gene is examined in more detail in section 4.3.3 on page 122.

In the case of the ENSG00000206551 gene, there was a ≈ 2.5 kb deletion in the chimpanzee orthologous region that affected the final non-coding area of the exon, but not the ORF, so the gene was retained (figure 4.3).

Initial methionine is required in eukaryotes

All eukaryotes initiate mRNA translation at an AUG codon (Kozak, 1999). Because of this we removed any genes for which none of the annotated proteins started with a methionine, as these are likely to be annotation artifacts.

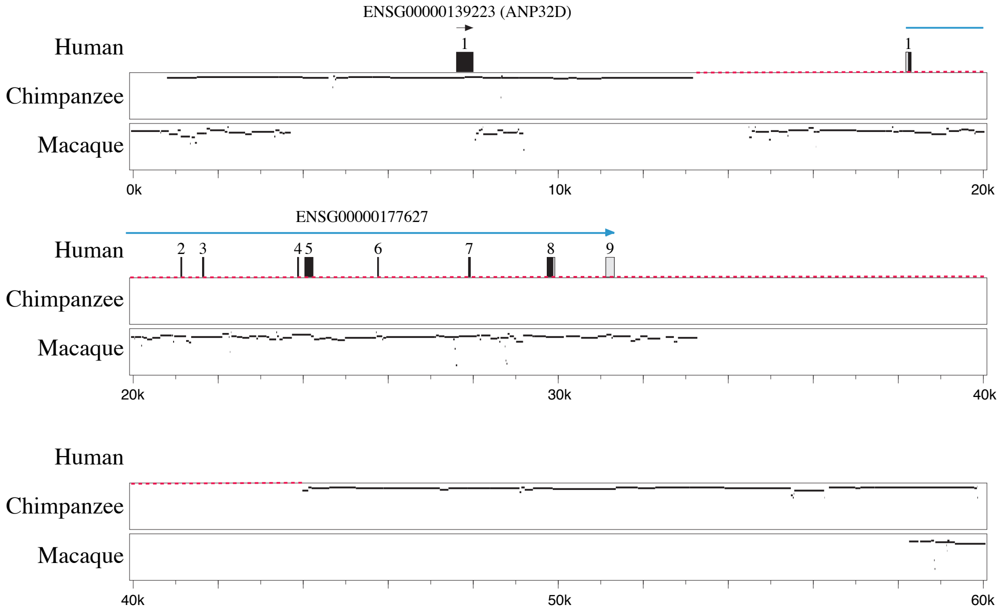


Figure 4.2: MultiPipMaker alignment between the human gene ENSG00000177627 and the orthologous region in chimpanzee and macaque. Boxes indicate human exons. The protein coding region is coloured in black and the UTR in grey. The lines within the two boxes marked as chimpanzee and macaque indicate the percent identity of the respective sequence when compared to human and ranges from 50% at the bottom of the box to 100% at the top. The exons belonging to this gene are marked with a blue arrow and the deletion in the chimpanzee genome that spans the complete area corresponding to the human gene ENSG00000177627 is marked by a red dotted line.

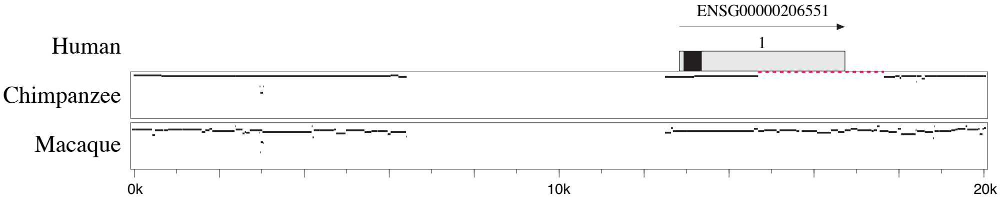


Figure 4.3: Alignment between the human gene ENSG00000206551 and the orthologous region in chimpanzee and macaque. The protein coding region is coloured in black and the UTR in grey. The lines within the two boxes marked as chimpanzee and macaque indicate the percent identity of the respective sequence when compared to human. This identity ranges from 50% at the bottom of the box to 100% at the top. The deletion in the chimpanzee genome is marked by a red dotted line.

This resulted in the removal of the ENSPTRG00000032478 chimpanzee gene for which the annotated protein was only 15 aa long, as well as two human genes ENSG00000183452 and ENSG00000198831.

The 24 remaining genes, all belong to human, contain a complete ORF, show sequence similarity to the syntenic area of chimpanzee and macaque and there were no deletions in the area corresponding to their ORF in either the chimpanzee or the macaque genomes.

Intron exon structure should be plausible

In many cases the automatic annotation system introduces small introns that prevent the occurrence of frame-shifts or stop codons in frame. There are many examples in which an annotated intron is certainly too short to be spliced being only a few nucleotides long. The shortest spliceosomal introns where evidence of splicing has been shown are 18 nt long (Gilson and McFadden, 1996) and the shortest intron reported for mammals, which were in humans, are 25 nt long (Deutsch and Long, 1999). Because of this we discarded any genes in which an annotated intron is found that is shorter than 18 nt and would introduce a stop codon or frame-shift if not removed by splicing.

The human gene ENSG00000198448 has a predicted five nucleotide intron between its two coding exons (exons two and three). This intron causes a frameshift that prevents a stop codon three aa after the predicted splice site. The gene is annotated as an EnsEMBL KNOWN gene and is supported by the UniProtKB/TrEMBL Q8N649 sequence which is the translation of an ORF contained in the BC027448 mRNA. However there is a discrepancy in the area where the intron has been predicted by EnsEMBL as the EnsEMBL sequence contains a TAG stop codon, and the sequence corresponding to the mRNA contains a TAT tyrosine codon. Because of this the gene was removed from the set.

Search for EnsEMBL orthologues removing those that are annotation artifacts

To examine the possibility of independent loss of these genes in both the chimpanzee and the macaque lineages we checked the EnsEMBL database for any genes that had been annotated as orthologues to any of the 23 remaining candidates in another species.

We found this to be the case for 14 out of the 23 remaining genes. Most of these putative orthologues were found in cat (*Felis catus*) and elephant (*Loxodonta africana*). The sequences of both of these species have a very low coverage, and many of the annotated genes have been identified by their similarity to genes described in other organisms. The presence of so many of the orthologues in these species and not in mouse (*Mus musculus*) which is a closer species with a higher quality genome sequence is surprising. Because of this we examined all the identified orthologues carefully. The presence of an orthologue in the database does not necessarily mean that orthologue is real, as it can easily reflect an annotation error.

In the process of exaptation a non-coding DNA fragment which will be present in the common ancestor of the two compared species becomes coding. This means that finding DNA sequence similarity at the syntenic location is expected. Automatic genome annotation will rely on the base composition of the area and the similarity to already annotated genes –mainly from human. Because of this areas that are similar to these genes, but present frame-shifts, in-frame stop codons or other features that make them non-functional may be predicted as genes, or in some cases as pseudogenes. Any in-frame stop codons and frame-shifts, may be compensated by the annotation machinery with the insertion of spurious introns that maintain the reading frame or skip the stop codons. These can be easily identified in most cases as the regions equivalent to one of the human exons will be broken into several smaller exons divided by introns that are below the minimum intron size.

The credibility of the orthologues predicted for each of these 14 candidates was examined in the same manner as was done previously for validating the candidates themselves.

We discarded any predicted orthologues that did not have an initial methionine in their encoded protein, and also removed any case that showed an unrealistic intron/exon structure, with introns under the minimum length described for eukaryotes. For 13 of these 14 candidate exaptations all predicted orthologues were found to contain unrealistic intron/exon structures and/or lack an initial methionine.

There was one remaining candidate gene, ENSG00000203917, this is an EnsEMBL Novel gene that is supported by a single cDNA from human. There is an orthologue annotated in the frog *Xenopus tropicalis*, although it is much shorter, and the similarity is not very high. The annotated protein shows a very low complexity with tyrosine and isoleucine forming more than 60% of the total protein, so it was discarded as a possible annotation artifact.

A summary of the results of this initial filtering can be seen in table 4.1 overleaf.

Table 4.1: Summary of the filtering process used for the 34 potential extinction/creation candidates.

Test	Genes removed
Assembly gaps affecting the coding region	6[a]
Deletion of the gene	1
Absence of an initial methionine in all the annotated proteins	3
Short introns that prevent a frame-shift	1
Annotated EnsEMBL orthologues	1[b]
Total removed	12
Remaining genes	22

[a] Another seven cases with small assembly gaps that did not affect the area aligned with the ORF were retained

[b] The gene ENSG00000203917 was removed, as the EnsEMBL orthologue was not clearly an annotation error, though gene itself was suspiciously low in complexity

4.3.2 Creation candidates

After removing 12 genes from the initial set of 34, the 22 remaining cases are good candidates for gene creation events by a process of exaptation of previously non-coding sequence. However, in order to be sure that they are not also expressed in chimpanzee or macaque we will determine if there is a conserved ORF in these species.

In 20 of them the area corresponding to the human gene was annotated as a pseudogene in either macaque or chimpanzee, in three of these cases it was annotated as a pseudogene in both. These were probably annotated as such based on similarity with human even if there is no evidence for a functioning ancestral gene.

When examining the results from a translated BLAT search (see chapter 3 on page 55) all of these 22 genes had a BLAT hit at the expected location in at least one of the other species and there were three cases where at least one in-frame stop codon was found by BLAT in the same position in both chimpanzee and macaque sequences. This makes parallel inactivation in both species unlikely, and these three genes would appear to be very clear examples of exaptation events in human. However, BLAT suffers from the same drawback as the automatic annotation system; it will adjust the reading frame of the hit in order to match the query sequence even when this requires the introduction of unrealistic introns. This means we may encounter the same problem as we found when examining the automatically annotated genes, where small frame-shift inducing gaps were introduced in order to maintain the reading frame used by the query. Because of this we cannot rely on the BLAT hits alone in order to decide if a gene is a clear example of exaptation, as there may be in-frame stop codons that are not found because the frame is altered by BLAT.

In order to detect if this occurred we examined in detail the genomic alignment of the three species for each protein coding area of these 22 genes. In this way we determined if any of the regions showed frame-shifts or in-frame stop codons in the area corresponding to the gene that would have been missed by our pipeline. In the automatic classification of the genes, in order to decide a gene was present but not annotated we required each of its exons to be present with an identity $\geq 90\%$ at the protein level and no in-frame stop codon (see section 3.5 on page 83). This threshold is very strict, and it is possible that in some cases the gene is conserved in one or both of the species but the identity is lower. The manual examination of the nucleotide level alignment will also identify any cases in which this may have occurred.

The gene ENSG00000184811is annotated as Tumor suppressor candidate 5 (TUSC5). When examining this gene we found in-frame stop codons in both the chimpanzee and the macaque syntenic regions.

This gene has been characterized by Oort *et al.* (2007) however when examining the protein sequence of the gene that is mentioned in this paper, we found that it did not correspond to the gene we had found. In order to determine if it was the same locus we used getorf from the EMBOSS package and determined if the mentioned protein sequence was indeed encoded by the gene annotated by EnsEMBL. We found that it was, and there was also a highly similar ORF encoded in the chimpanzee syntenic region, which means the gene is likely present in chimpanzee also.

We examined the newest version of EnsEMBL v49. We found that here this error had been corrected, and the correct protein was assigned to this gene. This means the gene is conserved in the three primates, so we removed it from the dataset.

Final classification of the remaining candidates

According to the genomic alignment we classified the remaining 21 genes as well as the ENSG00000177627 into three different categories:

- **Extinctions in chimpanzee:** In three cases the complete length of the ORF was conserved in macaque, however, the identity of the coding exons was lower than the threshold we had set in the automatic pipeline to remove cases in which the gene was conserved.

- **Uncertain:** In ten cases there is an uninterrupted ORF in either macaque, chimpanzee or both that is longer than half the human ORF. In some cases this ORF is not in the same reading frame as in human for its whole length, which means if it is expressed the resulting protein would be completely different, however because of the presence of a complete ORF we cannot exclude the possibility that it is also expressed in this other species.

- **Creation in human:** In 9 cases neither chimpanzee nor macaque have an uninterrupted ORF starting at the same point that is longer than half of the human predicted protein.

A summary of the genes classified into each category can be seen in table 4.2.

Table 4.2: Summary of the classification of the final human creation candidates.

Gene ID	Classification	Evidence[a]	Description
ENSG00000164621	Extinctions in chimpanzee	trans	Proline rich protein
ENSG00000180358		trans	-
ENSG00000177627		trans	-
ENSG00000175611	Uncertain	trans	-
ENSG00000177493		prot	Transmembrane protein
ENSG00000178554		trans	-
ENSG00000179421		trans	-
ENSG00000183633		trans	-
ENSG00000187229		trans	Nervous system abundant protein 11
ENSG00000196677		trans	-
ENSG00000197926		trans	-
ENSG00000204684		vega	-
ENSG00000206551		trans	-
ENSG00000178803	Creations in human	trans	-
ENSG00000196273		trans	-
ENSG00000204380		trans	-
ENSG00000204601		trans	-
ENSG00000204626		trans	-
ENSG00000205056		trans[b]	chronic lymphocytic leukemia up-regulated 1
ENSG00000205980		trans	-
ENSG00000206113		trans	-
ENSG00000206162		trans	-

[a] vega –vega/havana manually annotated transcript, trans –transcript level or prot –protein level.

[b] Buhl *et al.* (2006)

4.3.3 Gene extinctions in chimpanzee

We identified three cases of gene extinction in chimpanzee based on genome alignments. None of these cases had any useful annotation in the database, but they all showed evidence of transcription.

ENSG00000164621 (DAMS)

ENSG00000164621, also known as "10.3 kDa proline-rich protein DAMS" and also "SMAD5 opposite strand protein", is a proline rich protein located on human chromosome 5. As its name implies it is transcribed from the opposite strand to the SMAD5 gene. The first of this gene's two exons contains the complete coding region and is located within the first intron of SMAD5.

There is an intact ORF in the macaque genome, but the identity of the encoded protein is only 88% which is below the threshold used to determine whether a gene was present in the other organism. There seems to be a loss in the chimpanzee genome caused by a four nucleotide deletion from nucleotides 52 to 55 of the ORF. This produces a frame-shift resulting in the presence of a stop codon in-frame 13 aa later.

ENSG00000180358

ENSG00000180358 is located on chromosome 13, it has no description associated, but contains a serine phosphorylation site.

In the chimpanzee genome there was an open reading frame (which we define as a continuous stretch of nucleotides starting with a Methionine and ending in a stop codon) at the syntenic location. There is a stop codon in this chimpanzee ORF caused by an 8 bp insertion in the sequence of this species that produces a frame-shift and a premature stop codon although the existing ORF is longer than half of the length of the ORF in human. In the macaque, the gene seems to be present as the ORF is conserved except for a 6 bp deletion (or insertion in the hominoid lineage) that does not alter the reading frame.

ENSG00000177627

ENSG00000177627 is located on chromosome 12 and has no annotation associated to it. It is also the only remaining candidate that presents a complex intron-exon structure with the ORF spanning more than one exon, which makes it a less likely candidate for *de novo* generation of a coding gene.

The reason why this gene had not been removed by the automatic filtering and classified correctly was because exons 2 and 3 are quite short and not identical to the human exons, so although they are present in macaque they had been missed by the BLAT search. Because all exons were not found and not all of the identified ones had an identity $\geq 90\%$ the possible macaque orthologue had been discarded. However, this gene is likely to be still present in the macaque, as the five exons found by BLAT conserve a high identity ($> 84\%$) and there are no frame-shifts or in-frame stop codons. Thus it appears to be a case of gene extinction in the chimpanzee lineage.

4.3.4 Creation of new human genes by exaptation of non-coding sequence

We found nine human genes where the genomic sequence of the chimpanzee and macaque were highly similar at the expected location, but contained no ORF covering at least half of the human gene (table 4.3 overleaf). Five of these genes show an overlap with at least one other gene, and in all cases they are close to a CpG rich area. In humans and mice 60% of all promoters co-localize with C+G rich regions devoid of methylation, which are known as CpG islands (Antequera, 2003). A brief summary of those cases that appeared more interesting is shown.

ENSG00000204601

The human gene ENSG00000204601 is located on chromosome 12, and has evidence of expression in brain, heart and lung tissue ((Q4G0G9 mRNA) as well

Table 4.3: Final gene creations.

Gene ID[a]	Expression[b]	Overlapping gene[a]	CpG[c]
ENSG00000178803 −	Kidney	ENSG00000128271 (Adenosine receptor A2a) +	Yes
ENSG00000196273 +	Placenta	-	Yes
ENSG00000204380 −	Trachea	ENSG00000144283 (Plakophilin-4) +	Yes
ENSG00000204601 +	Brain, heart, lung and testis	-	Yes
ENSG00000204626 −	Hyppocampus	ENSG00000197653 (Dynein, axonemal, heavy polypeptide 10 isoform 1) +	Yes
ENSG00000205056[d] +	blood	ENSG00000205057 (CLLU1OS) −	Yes
ENSG00000205980 −	testis	ENSG00000132405 (TBC1 domain family member 14) +	Yes
		ENSG00000173011 +	
		ENSG00000173013 (Coiled-coil domain-containing protein 96) −	
ENSG00000206113 +	testis	-	Yes
ENSG00000206162 −	cerebellum	-	Yes

[a] The strand from which the gene is transcribed is indicated by a +/− sign

[b] Tissues in which there is evidence of expression, usually in the form of mRNA

[c] Area with a CG content $\geq 60\%$ within 10 kb of the gene

[d] Buhl *et al.* (2006)

as testis (AK098523 cDNA). The transcript is formed by three exons, and the protein is coded for by exon 3 alone. The possible ORFs in chimpanzee and macaque both end at the same position at a stop codon 41 aa after the initial methionine. An A inserted at position 10 of the human genome eliminates the downstream stop codon in this species allowing for a longer ORF. The parallel inactivation of this gene in both chimpanzee and macaque is extremely unlikely as it would have required the same mutation in both species. The accuracy of the human, chimpanzee and macaque sequencing is supported by the sequence traces, which show the presence of two adenine residues in the human lineage and only one in both the chimpanzee and the macaque (figure 4.4 overleaf).

CLLU1

CLLU1 (ENSG00000205065) is located on human chromosome 12. It has two orthologues annotated in EnsEMBL one in the guinea pig (*Cavia porcellus*) and one in the tree shrew (*Tupaia belangeri*), however in both of these species a close examination of the annotated sequence revealed they are unlikely to be real due to the absence of an initial Met in the case of the guinea pig, and the insertion of several spurious short introns that prevented frame-shifts in the case of the tree shrew.

This gene was experimentally identified when searching for genes that are differentially expressed in chronic lymphocytic leukemia (CLL) by Buhl *et al.* (2006). CLLU1 transcript levels were only detected at appreciable level in CLL cells, but not in any normal tissue or tissue from other hematologic malignancies. The region containing the gene is very dense in ESTs derived from germinal B cell and CLL cells which suggests the chromatin in this area possesses an open structure easily accessible to transcription factors in B cells (Buhl *et al.*, 2006). The CLLU1 gene encodes 6 mRNAs of which only two potentially encode a peptide, all of these mRNA were shown to be expressed. No known miRNAs were identified in these transcripts and the necessary hairpin structures required

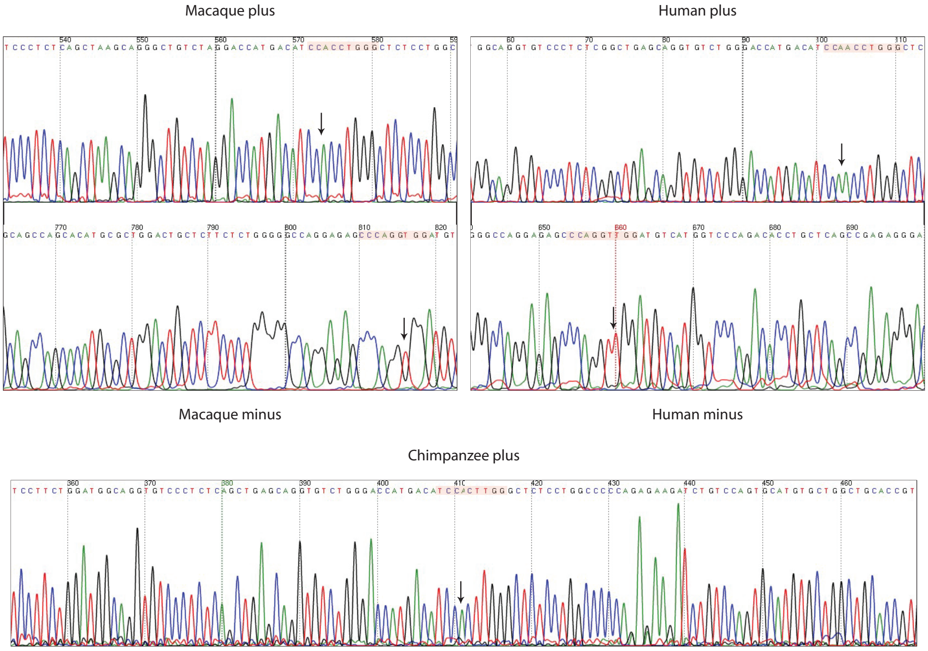


Figure 4.4: Gene sequence traces from the region corresponding to the ENSG00000204601 human specific mutation and the corresponding areas in chimpanzee and macaque. For human and macaque the best hitting trace is shown for both the forward and reverse strand. In the case of chimpanzee there was only a trace sequence trace from the forward strand, however the sequence is unambiguous. The region where the difference between the three species occurs is highlighted and marked with an arrow.

for the generation of miRNA were not detected either. The inferred peptide encoded shows a remarkable structural similarity to human interleukin 4 (IL-4), which suggests the possibility of this peptide activating the IL-4 pathway that would decrease the sensitivity of CLL cells to apoptotic stimuli (Buhl *et al.*, 2006).

When examining the alignment of the syntenic area in macaque and chimpanzee we found a high degree of conservation in both with no sequencing gap in either chimpanzee or macaque in the whole length of the area aligning with the human gene (figure 4.5).

We examined the area corresponding to the annotated CLLU1 ORF in detail and found there has been a deletion of an A in position 123 in the human sequence, that is present in both chimpanzee and macaque. The presence of this extra A causes a frame-shift in these two organisms with respect to human that introduces a stop codon in-frame one amino acid later (figure 4.7). In this case the parallel occurrence of a the same mutation in both macaque and chimpanzee seems very unlikely. The sequence traces for the three species were examined and they support the presence of an extra *A* in the chimpanzee and macaque genomes, but not in the human (figure 4.6 overleaf).

Notably this gene is overlapping another annotated gene, CLLU1OS (CLLU1 opposite Strand) which is transcribed form the minus strand in the same area. This transcript was also identified in the study by Buhl *et al.* (2006) in which it was named cDNA7. We identified a possible one-to-one orthologue to the CLLUOS1 gene in chimpanzee, which has the first two coding exons conserved, but shows a premature stop codon in the middle of the third coding exon. The other orthologues annotated by EnsEMBL for this gene show unlikely intron-exon structures.

We suggest from the information available that once the deletion occurred in the human lineage an ORF of sufficient length was available in an area that already presented an open chromatin conformation. This would facilitate the

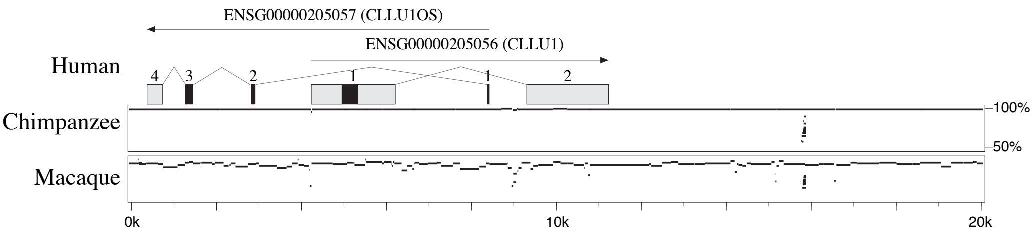


Figure 4.5: Alignment between the human gene CLLU1 and the orthologous region in chimpanzee and macaque. The protein coding region is coloured in black and the UTR in grey. The lines within the two boxes marked as chimpanzee and macaque indicate the percent identity of the respective sequence when compared to human. This identity ranges from 50% at the bottom of the box to 100% at the top. The gene named CLLU1OS by EnsEMBL corresponds to a transcript from the same region that originates from the reverse strand.

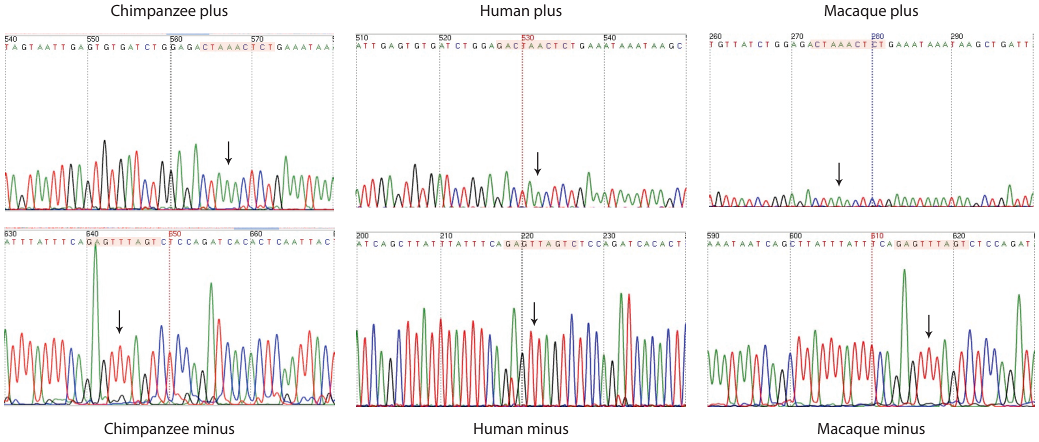


Figure 4.6: Gene sequence traces from the region corresponding to the CLLU1 human specific deletion and the corresponding areas in chimpanzee and macaque. The best hitting trace is shown for both the forward and reverse strand. The region where the difference between the three species occurs is highlighted and marked with an arrow.

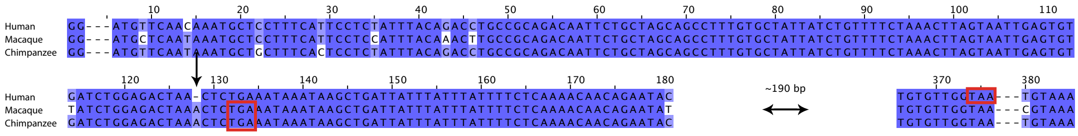

10 20 30 40 50 60 70 80 90 100 110
Human GG---ATGTTCAACAAATGCTCCTTTCATTCCTCTATTTACAGACCTGCCGCAGACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Macaque GG---ATGCTCAATAAATGCTCCTTTCATTCCTCCATTTACAAACTTGCCGCAGACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Chimpanzee GG---ATGTTCAATAAATGCTGCTTTCACTCCTCTATTTACAGACCTGCCGCAGACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
120 130 140 150 160 170 180
~190 bp
370 380
Human GATCTGGAGACTAA-CTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAACAGAATAC TGTGTTGGTAA---TGTAAA
Macaque TATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAACAGAATAT TGTGTTGGTAA---CGTAAA
Chimpanzee GATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAACAGAATAC TGTGTTGGTAA---TGTAAA

Figure 4.7: Detail from the nucleotide alignment between the human gene CLLU1 and the orthologous region in chimpanzee and macaque. The black arrow indicates the human specific deletion that allows for the longer ORF, and the red box indicates the stop codons that would terminate transcription in each of the species.

expression of any gene within the area even if the promoter signals were weak, as would be the case for a potential case of *de novo* generation of a gene by exaptation of a non-coding region. This is also confirmed by the high density of the EST coverage in the region.

This gene was first detected as expressed in cancer cells, and it is unclear what its normal function may be, or if it is indeed expressed under normal circumstances.

4.4 Discussion

Most of the cases of new genes that have been reported can be traced back to duplication events either of complete or partial genes. In this chapter we examine the possibility of *de novo* generation of genes from non-coding sequence based on the fact that many redundant ORFs exist distributed throughout the genome. Though improbable it is possible that one of them may acquire the necessary signals for its transcription and subsequent translation. Indeed there have been some recent reports of new genes that have apparently originated *de novo* from non-coding DNA both in *Drosophila* species (Levine *et al.* 2006, Begun *et al.* 2007) and in *Saccharomyces cerevisiae* (Cai *et al.*, 2008).

Many of these ORFs may have originated by degeneration of previously functional genes that originated by gene duplication. The generation of duplicates, as has been noted by Lynch and Conery (2000) is far more frequent than has been thought before, and because in most cases these genes will be rendered inactive by deleterious mutations we would expect the remnants of many genes to be distributed along the genome for a period of time until they degenerate beyond recognition, as can be observed from the abundance of pseudogenes present in the genome of most mammals. This means that if an ORF is recruited from this pool of dead genes it may show some similarity to known genes and could potentially contain certain domains that are still functional (Brosius, 1999).

De novo expression of an open reading frame that spans more than one exon would be less likely than those ORFs that are confined within an exon, as the former would require the presence of splice sites that can maintain the reading frame through the exons, as well as the promoter signals. Indeed, in all the cases of creation candidates, the annotated ORF is confined to one single exon. Although the transcripts possess more exons (2 to 4 in the nine final candidates), the ORF is contained entirely within one of them. In most cases this coding exon is quite large, with the ORF occupying less than half of the total sequence of the exon.

From the initial 34 cases all those genes that contained more than one coding exon were removed because of assembly gaps in the areas corresponding to the coding region, except for the case of ENSG00000177627 which was caused by the deletion of this gene in chimpanzee.

An ORF may conceivably acquire the necessary promoter signals in several ways. Just as the coding areas of a gene can become duplicated and translocated to a different region in the genome, promoter regions may undergo the same process. If this occurs, and the sequence is translocated close to an existing ORF it may induce its expression.

Another way in which an ORF can obtain the necessary promoter signals is the recruitment of the promoter region from another nearby gene. Unlike prokaryote promoters some elements of the eukaryote promoters such as the CAAT box and the GC box can function when located on the opposite strand of the gene being expressed, also the distance at which certain elements such as enhancers need to be from the initiation of transcription is more flexible with some eukaryotic enhancers being able to act over distances of several kb (Berg *et al.*, 2002). Because of this some ORFs may be able to use these elements from another gene.

When examining the nine final creation candidates that passed all our filters, five of them (ENSG00000178803, ENSG00000204380, ENSG00000204626, ENSG00000205980 and CLLU1) were overlapping at least one other annotated human gene that was transcribed from the opposite strand. Another three had an annotated gene within ten kb upstream from their first exon. In the case of CLLU1 it was overlapping the gene annotated as CLLU1OS, so we did not count this as an interesting overlap because there is not stronger evidence for the overlapping gene than for the exaptation candidate.

It has been observed that the fraction of overlapping genes with orthologues in other organisms is significantly lower than the total fraction of genes with orthologues in the same related species (Makalowska *et al.*, 2007). This may suggest that many overlapping genes are new additions to the genome, created by the exaptation of a sequence within or very close to an existing gene (Makalowska *et al.*, 2007). As mentioned above a new gene created in this manner could use the already existing transcription signals of the pre-existing gene, which means it may not require a *de novo* creation of promoter and enhancer signals in order to be expressed.

Previous studies have shown that 60% of the mammalian promoters are associated with CpG islands (Antequera, 2003). Although not all GC rich areas necessarily correspond to promoters, the presence of a GC rich area close to the first exon may be a good sign. When examining these nine candidates all of them had GC rich areas within 10 kb of the first annotated exon with a GC content $\geq 65\%$.

In the only case in which the examined gene has been studied in detail, that of CLLU1, we found that indeed it is contained within a region that facilitates its expression in the cell type within which it is expressed, probably due to an open configuration of the chromatin that facilitates the access of transcription factors. Although the reason why the gene is highly over-expressed in CLL cells is not clear, the expression appears to cause a deleterious effect, which we may expect if a random gene is suddenly expressed at a high level. In this particular case as it shows certain structural similarity to IL-4, this fact, and the features of the cells where it is over-expressed seem to indicate the mechanism of interference is through the inhibition of the apoptosis pathway (Buhl *et al.*, 2006).

The fact that the expression of this gene seems to be correlated with a negative effect in the cells where it is expressed would not be unexpected, as it may interfere with the already existing processes within the cell. However, as the high expression level is not constitutive, the presence of this new protein may not cause any deleterious effect in normal circumstances.

Interestingly three out of the nine candidates are expressed in testis. In this tissue, during the haploid stage of spermatogenesis the transcription environment is very permissive, with a large increase in the amount of polymerase II enzymatic complex available (Schmidt and Schibler, 1995). This larger amount of available polymerase may allow for the expression of regions in the genome that possess suboptimal promoters. Overlapping genes may have a similar effect, as during their own transcription the chromatin in the area will be in an open conformation and the polymerase complex will be recruited to the area. This may also allow for any suboptimal promoters present in the area to be used for transcription initiation.

The other exaptation candidates will require more evidence in order to determine if they indeed possess any function or affect any cellular processes.

Chapter 5

Duplication and Subfunctionalization of alternatively spliced genes in *Homo sapiens* and *Pan troglodytes*

5.1 Introduction

Gene duplication is an important source of new genes (Ohno 1970, Nei and Rooney 2005, Katju and Lynch 2006). Whole genome duplication, segmental duplication and tandem duplication of areas that contain genes are the most common mechanisms of gene duplication that produce functional duplicates. Both of these mechanisms allow the duplication of the elements surrounding the duplicated gene that are necessary in order to obtain two transcriptionally active genes that are identical both in sequence and expression. The importance of gene duplication in the evolution of vertebrates has been noted many times since the original suggestion by Ohno (1970). Several recent large studies have shown the important role duplication events, from single gene duplication to whole genome duplication, have had during evolution. These studies range from differences between human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*) observed by Cheng *et al.* (2005), to those between different vertebrate species (Blomme *et al.*, 2006), or between more distant eukaryotes (Maere *et al.*, 2005).

After gene duplication different evolutionary fates await the resulting duplicates (Lynch and Conery 2000, Prince and Pickett 2002). In the classical model, these genes had two possible fates: the loss of one of the duplicated members through accumulation of degenerative mutations, or preservation of both copies due to beneficial mutations that confer an advantage to the organism in one of the copies. A third outcome was added by the Duplication-Degeneration-Complementation (DDC) model (Force *et al.*, 1999) also called subfunctionalization.

The subfunctionalization model proposes that the absence of selective pressure caused by the redundancy of the duplicated genes allows each of the copies to accumulate mutations as long as the combination of both genes maintains the original functions. Most genes have more than one function. and each of these functions may be controlled by different elements within the gene. These elements are directly responsible for the particular biochemical reaction, cellular localization of a gene product and expression levels, tissues and times required for the correct performance of each function. These different elements may be located in different areas of the gene and surrounding DNA, and each of them can be affected independently by mutations. If a mutation affects a specific function of the ancestral gene in one of the copies and the other copy loses a different function by mutation in a different area, this may lead to the retention of both copies of the gene in order to maintain the full set of ancestral functions (Force *et al.*, 1999). This type of complementary mutation occurring on both duplicates prevents any one of them from fulfilling all the original functions of the pre-duplication gene on its own, but maintains the original set of functions by dividing them between the resulting duplicates (figure 5.1).

Immediately after a gene duplication if the resulting copies are identical, mutations will start to accumulate on both. It is likely that each copy will suffer mutations that will affect different functions, and when one function is completely lost from one of the duplicates, it will be fixed on the other by the selective pressure to maintain the complete set of functions. This process will cause the retention of duplicated genes without the need for any advantageous mutations. The number of mutations that may affect one of the gene functions is larger if the gene has many functions, and in this case it is also more likely that mutations will affect different functions in each of the copies. Because of this, the subfunctionalization process will be more likely to occur the more functionally complex the original gene was (Force *et al.*, 1999). The loss of one duplicate may be considered as a special case of this in which one of the copies loses all the functions and the other retains them, and would be more likely in the case of genes with few functions.

The division of the functions between the two resulting duplicates will allow further independent specialization of each of the copies in the specific function they retained. This is particularly advantageous in cases where specialization in one of the functions by the ancestral gene would have affected the other in a negative way (Hittinger and Carroll, 2007).

Alternative splicing allows the production of different products from a single gene, and plays an important role in the large complexity of eukaryotes. More than 40% of vertebrate genes may undergo alternative splicing, and this has been suggested as one of the reasons for the higher complexity of mammals when compared to invertebrate organisms whose gene content is not that different (Kim *et al.*, 2007).

Genes that have alternative splice variants may be duplicated, and because of their particular features they are ideally suited for subfunctionalization by differential loss of splice variants. This is because the regions of the gene that

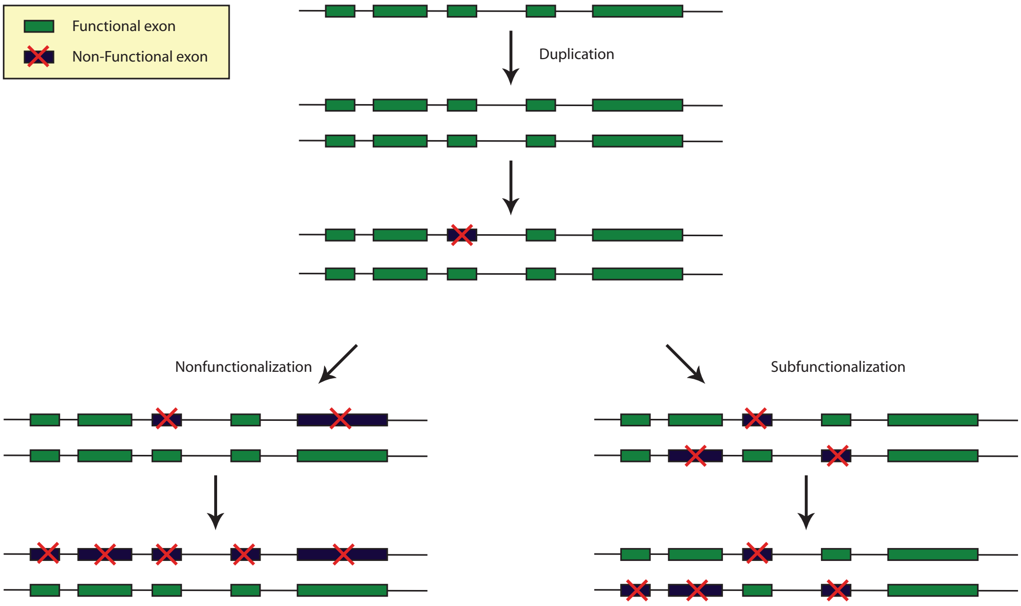


Figure 5.1: This figure shows two of the potential fates of duplicated genes. The small green boxes show different exons each of them involved in a particular function of the original gene. However any kind of functionally discreet and independent locus (such as protein domains or regulatory elements) could be affected by the subfunctionalization process. After duplication, mutations start to accumulate some of which may prevent the correct splicing of certain exons in one of the copies or damage important functional areas within them. On the left these mutations continue to accumulate in one of the duplicates leading to the complete inactivation of this copy. In the right side degenerative mutations occur in complementary exons in the two copies, this way maintaining between the two duplicates the complete set of splice forms and so all the functions of the original gene. If all these functions are essential none of the copies can now be lost (Force *et al.*, 1999).

are specific to each of the alternative forms, being different exons, are already separated in a way that allows the different forms to be affected independently by deleterious mutations. Even in the case of overlapping alternative exon in different reading frames, mutations such as stop codons can be introduced in ways that will affect one reading frame and not the other.

There have been cases of this particular type of subfunctionalization described in different organisms.

The microphtalmia-associated transcription factor (*Mitf*) has different isoforms in birds and mammals that are generated through the use of alternative 5′ promoters. This gene plays a role in differentiation and survival of melanocytes. In teleost fish species two separate genes (*Mitf*-m and *Mitf*-b) exist. Each of these genes encodes a protein that corresponds to one of the bird/mammalian isoforms, and the two of them have different expression profiles. Degeneration of the first exon which is present in the mammalian MITF-m form is observed in the *Mitf*-b gene in fish but not in the fish *Mitf*-m form (Altschmied *et al.*, 2002).

Human encodes three synapsin genes (*Syn1-3*), in the pufferfish (*Takifugu rubripes*) there is a second copy of *Syn2*, (*TrSyn2B*). In human *Syn2* generates two alternatively spliced variants, but in the pufferfish each of these variants is encoded by one of the two different *TrSyn2* genes, and both of these duplicated genes have lost the ability to produce the form encoded by the other duplicate through the accumulation of complementary degenerative mutations (Yu *et al.*, 2003).

In plants the chloroplast gene *Rpl32* was relocated into the nuclear genome and fused with a superoxide dismutase genes (*Sodcp*) some time before the divergence of mangrove (*Bruguiera gymnorrhyza*) and poplar (*Populus trichocarpa*) forming the *Sodcp-Rpl32* gene. This gene is alternatively spliced in mangrove but became duplicated in poplar. Each of the resulting duplicates encodes a single protein that corresponds to one of the alternative forms of the mangrove *Sodcp-Rpl32*, and due to degenerative mutations has lost the ability to produce the other form (Cusack and Wolfe, 2007).

In these subfunctionalization examples the resulting duplicates have fewer alternative forms. This is something that has been observed for duplicated genes in general. When compared to single copy genes, duplicated genes and genes belonging to large families have fewer alternative forms, an observation that supports the subfunctionalization model of genes with alternative forms early after the duplication event (Su *et al.*, 2006).

To our knowledge there have been no comprehensive searches for possible cases of duplicated genes with alternative splice forms that may have originated and undergone subfunctionalization since the divergence of the human and chimpanzee lineages. The availability of the draft genomic sequence of the chimpanzee (*Pan troglodytes*) and the complete sequence of the human (*Homo sapiens*) genome allow us to search for any case of duplication with posterior subfunctionalization that may have occurred between the two species since the divergence of both lineages 6 Mya, and might have contributed to the differences observed between the extant chimpanzee and human species.

5.2 Materials and Methods

5.2.1 BLAST search

An all-against-all BLASTp search was done using all the alternative protein forms encoded by each of the genes in the genomes of *H. sapiens* and *P. troglodytes.* Protein hits within a species were ignored (see section 3.2.2 on page 60).

5.2.2 Alignments and phylogenetic tree construction

Alignments were obtained using t-coffee with the default parameters with no penalty for terminal gaps.

Neighbor joining trees were built with ClustalW (Thompson *et al.*, 1994) with a bootstrap value of 1000, using Kimura's correction for multiple hits and ignoring positions with gaps unless otherwise stated.

Alignments and trees were evaluated manually.

5.2.3 Domain structure

The domain structure of the different proteins was examined using SMART (`http://smart.embl-heidelberg.de/`, Schultz *et al.* 1998).

5.2.4 EST support

In order to determine EST support the transcript sequences were searched against the EST database. The search was done using MegaBlast which is optimized for finding sequences with high similarity, with an E-value cutoff of $1e^{-4}$, no masking and a fixed database size of $1e^{9}$, in order to obtain E-values comparable to the other searches we have done (see chapter 2 on page 51). We selected those EST hits annotated from the same species, in which the identity level of the hit was ≥ 0.95 and the hit contained a stretch of 100 or more identical residues.

5.3 Results

5.3.1 Search for initial candidates

If a gene that undergoes alternative splicing has been recently duplicated since the divergence of the human and chimpanzee we expect the alternative transcripts of the unduplicated gene to match transcripts from different genes (*i.e.* the two daughter genes) in the other species.

We searched for evidence of this by performing an all-against-all BLAST search using all the proteins annotated for each of the genes and extracted those cases in which one gene (A) encoded two or more proteins(a_1,a_2) that each had as a reciprocal top hit (RTH) a protein ($a_1 \Leftrightarrow b_1$ and $a_2 \Leftrightarrow b'_1$) from two different genes (B, B'). In this initial step we obtained 264 candidate genes that were alternatively spliced and for which different gene products hit proteins encoded by 914 different genes as RTHs. 141 of these belonged to human and 123 to chimpanzee.

5.3.2 Candidate filtering

Because duplication with subsequent subfunctionalization is not the only reason why an alternatively spliced gene may have different RTHs we devised a series of filters in order to remove those cases which were likely to be caused by other reasons. The following cases were removed:

- If genes B or B' had another almost equally good BLAST hit (within an e-value threshold of $1e \times 10^3$ from the top e-value) different from the alternatively spliced gene A, the gene was removed from the set as this indicates a many-to-many relation between the genes and not a one-to-many (particularly one-to-two) relation which is what we are searching for. This occurred in 812 of the initial query-hit combinations that were obtained from the BLAST search. This may happen if both genes A and B are alternatively spliced genes that were duplicated recently, as both would hit each other or if both belonged to a large family of genes that have high sequence similarity.

- We removed those cases in which one of the proteins of the alternatively spliced gene (A) hits proteins belonging to more than one gene in the opposite species. This occurred in five cases. The most likely cause for this is the presence of common domains that are conserved between several genes.

- We examined the candidate alternatively spliced genes (A) and removed those gene hits (B) resulting from transcripts from gene A that did not overlap any other transcript from the same gene (A). If the transcripts do not overlap in any way, the alternative forms could be annotation artifacts where two different genes that are adjacent or located close by in the genome have been incorrectly annotated as a single one. This occurred in 44 cases, an example of it can be seen in the ENSG00000092529 gene. This gene encodes two different known proteins, calpain 3 (CAPN3) and glucosidaseα neutral C (GANC). Two of the transcripts annotated for the CAPN3 region overlap on their non-coding region with the transcript encoding GANC, and because of this the two genes have been annotated as one by the automatic annotation pipeline (see figure 5.2 overleaf) These two transcripts have truncated versions of the protein due to changes in the reading-frame, and were not part of the transcripts from ENSG0000095529 that had different genes as reciprocal best hits, because of this the two sets of remaining transcripts for the gene did not overlap and were removed by the pipeline.

- We examined both the alternatively spliced candidates an the genes they hit for cases that may be uncertain and 294 genes that did not have an EnsEMBL KNOWN status were also removed. This is done in order to avoid false positives that may be annotated by the gene prediction algorithms but have no further evidence.

- We removed all cases in which the protein encoded by one of the duplicated genes (B) did not start with an initial methionine. Eukaryote translation always starts with an AUG codon (Kozak, 1999). Proteins that do not start with a Met are likely to be annotation or sequencing errors where part of the protein is missing, or they could be caused by a gene that is genuinely truncated and is unlikely to be functional. This occurred in 82 cases.

.

After removing those genes which showed one or more of these problems there were 14 genes that were present as a single gene in one of the species but were

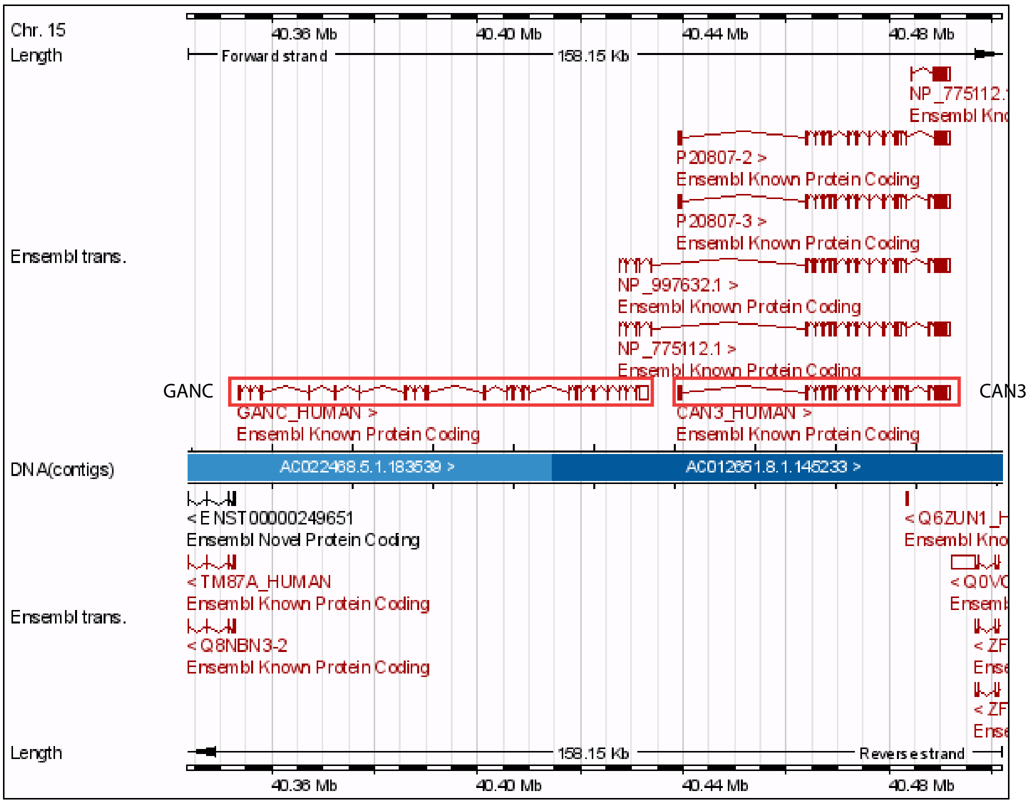


Figure 5.2: Schematic representation of the exon structure of the EnsEMBL ENSG00000092529 gene which appears to be two separate genes, GANC and CPN3, that have been misannotated as one. The separate genes are indicated by red boxes. This figure was obtained from the EnsEMBL genome browser

potentially duplicated in the other. Two of them were human genes and twelve belonged to chimpanzee.

5.3.3 Final Candidate Filtering

For each of the 14 cases the exon structure of the alternative forms was plotted against the genomic sequence and examined for any indication that the gene may be incorrectly annotated. The analysis of each of the 14 candidates is summarized in table 5.1 on page 145. We found four cases in which none of the coding exons of the two transcripts overlapped (an example can be see in figure 5.3 overleaf) and six in which there are two separate groups of overlapping alternative transcripts that do not overlap or share any coding exons with transcripts of the other group (figure 5.4 overleaf). In all of these cases the different sets of non-overlapping transcripts hit different genes and this is the reason why they were selected as potential candidates.

Six of these 14 cases appear to be annotation errors in which two highly similar genes that appear to be recent tandem duplications have been incorrectly merged into one in the annotation. We investigated the reason why these genes had not been removed by our automatic pipeline check for non-overlapping transcripts and we found in all cases these transcripts that were kept were kept because they overlapped with another of the transcripts of the same gene, although not all of these remaining groups of transcripts overlap with each other as is the case of the SEMG2 gene (figure 5.4). In the other four cases the two groups of transcripts code for different products but their transcripts also overlap (figure 5.3). These ten cases which included both of the human genes were removed.

In order to verify these suspected annotation errors we also examined the locations in the other species genome of the putative duplicated genes. Indeed in all the cases except for the KLRC4 these genes were adjacent in the genome of the other species or separated by only one gene, so they were removed from the dataset. In the case of KLRC4 hits one of these genes was located in an

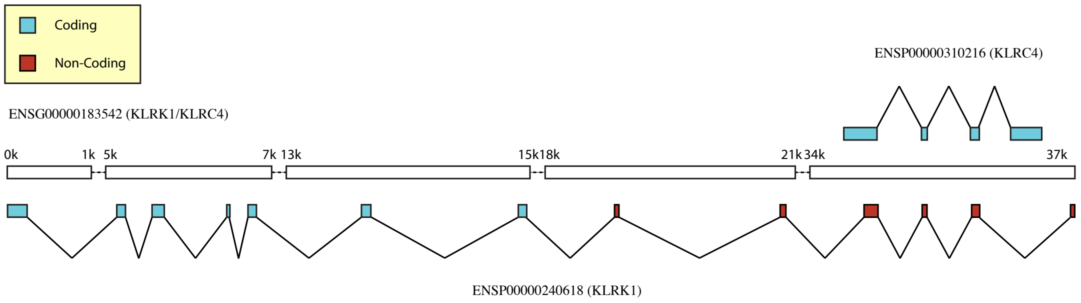


Figure 5.3: Schematic representation of the exon structure of the two candidate alternative transcripts belonging to the human gene ENSG00000183542 the first transcript is annotated as Killer cell lectin-like receptor subfamily K member 1 (KLRK1) and the second transcript as killer cell lectin-like receptor subfamily C, member 4 (KLRC4). The coding exons of these transcripts show no overlap. The genomic DNA sequence is indicated by the white rectangles, with some regions which do not contain exons shortened and indicated by a dotted line. The transcripts are shown aligned to the genomic sequence with those exons that are non-coding in red and those that are complete or partially coding in blue.

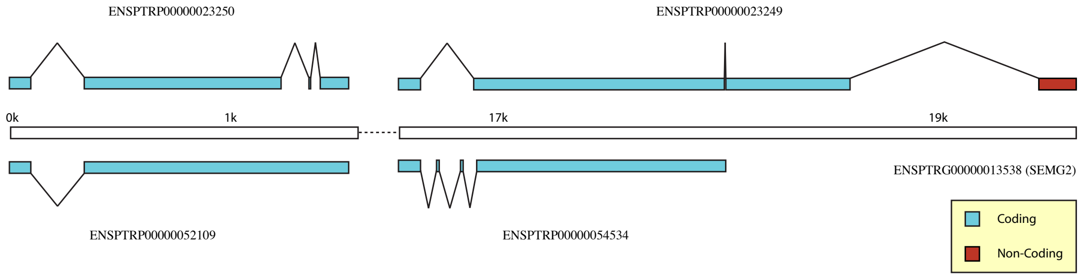


Figure 5.4: Schematic representation of the exon structure of the four candidate alternative transcripts belonging to the chimpanzee SEMG2 gene. Each pair of overlapping transcripts corresponds to a different gene in human, and these two human genes are consecutive on the human genome genome. The genomic DNA sequence is indicated by the white rectangles, with some regions which do not contain exons shortened and indicated by a dotted line. The transcripts are shown aligned to the genomic sequence with those exons that are non-coding in red and those that are complete or partially coding in blue

Table 5.1: Summary of the candidate genes, and those genes thay hit in the other species, as well as the common name and description, if available, and the final classification of the gene.

EnsEMBL_gene_id[a]				
Gene *A*	Gene B/B'	Common name[b]	Description[b]	Classification
ENSG-183542	ENSPTRG-004672 / 032649	KLRK1 / KLRC4	NK cell receptor D	2 genes merged
ENSG-196565	ENSPTRG-028731 / 022526	HBE1 / HBG1 / HBG2	Hemoglobin subunit $\gamma 2$	3 genes merged
ENSPTRG-003121	ENSG-174775 / 185522	-	-	3 genes merged
ENSPTRG-009528	ENSG-189162 / 136487/204414	SOM2	Growth hormone 2	3 genes merged
ENSPTRG-013538	ENSG-124233 / 124157	SEMG2	Semenogelin 2	2 genes merged
ENSPTRG-014457	ENSG-189306 / 182841	-	-	2 genes merged
ENSPTRG-016478	ENSG-153143 / 170180	GLPA	Glycophorin A	2 genes merged
ENSPTRG-019449	ENSG-013455 / 130429	-	Predicted Actin related protein 2/3 complex subunit 1A	2 genes merged
ENSPTRG-019780	ENSG-127364 / 127366	TAS2R5	Taste receptor	Truncated alternative form
ENSPTRG-019844	ENSG-133624 / 181220	-	-	2 genes merged
ENSPTRG-019953	ENSG-164821 / 164822	DEF1	Neutrophil defensin 4	2 overlapping genes
ENSPTRG-019963	ENSG-178287 / 164871	SPG11	Sperm associated antigen 11	Subfunctionalization
ENSPTRG-021633	ENSG-099725 / 183943	PRKY	-	Subfunctionalization
ENSPTRG-021919	ENSG-126752 / 171483	SSX	Predicted similar to synovial sarcoma, X breakpoint 6	Subfunctionalization

[a] The gene names used are the EnsEMBL identifiers, but for space purposes a '$-$' corresponds to '00000'. Those genes starting with 'ENSG' belong to human and those starting with 'ENSPTRG' to chimpanzee.

[b] Common name and description are taken from gene*A*.

unassembled fragment of the same chromosome as the other (chromosome 12 random), so this case was also discarded as dubious.

After eliminating the cases with no coding exon overlap there were only four candidates remaining (TAS2R5, SPG11, PRKY and SSX) we examined the alignment of each alternative form with the corresponding proteins of the other species. If the case is a recent subfunctionalization we expect the alignment between the alternative transcript and the new subfunctionalized protein to span the complete length of the protein, and each of the alternative forms to align with one of the subfunctionalization candidates. If this is not the case it could indicate either a partial duplication if the new gene is smaller or a more complex duplication history if it is longer.

In the case of the chimpanzee TAS2R5 (PtTAS2R5) this gene is located on chromosome 7 in chimpanzee and we classified it as an orthologue to the human TAS2R5 (HsTAS2R5) gene which is located also on chromosome 7. It belongs to a family of taste receptors involved in the perception of bitter taste. These genes in human are organized in clusters in chromosome 7 and 12 which also contain many pseudogenes. This gene is located within one of the human-chimpanzee synteny blocks and when examining the genes surrounding it we find the human gene TAS2R4 is adjacent on the genome and does not have an orthologue annotated on chimpanzee, while the other genes in the vicinity all have one-to-one orthologues (figure 5.5 on page 148). In addition to this, the two genes which are hit by PtTAS2R5 are HsTAS2R5 and HsTAS2R4, however the transcript that hits HsTAS2R4 is truncated and missing part of the sequence. In order to determine if this is not also a false positive that had not been removed because of the overlap between the two transcripts we examined the gene in more detail. We aligned the genomic region from the two species (figure 5.6 on page 148) and found that the region corresponding to the HsTAS2R4 contains a 0.5 kb in-del by which the centre of the coding area of TAS2R4 is missing in chimpanzee although the rest of the region is highly identical, so if it is a duplication it occurred before the two lineages diverged. We examined the splice variants annotated by EnsEMBL that had similarity to HsTAS2R4. There were 6 transcripts, all of them annotated as novel that started in the region corresponding to TAS2R4 and were then spliced to a region corresponding to the TAS2R5 gene. All of these taste receptors are encoded in human by a gene that possesses a single exon, while this chimpanzee gene is annotated as having not only several exons, but also 8 different splice variants, 6 of them forming a final protein that appears to be a hybrid between TAS2R4 and TAS2R5. Both the genome alignment data and the presence of these hybrid transcripts appear to indicate a case of misannotation, by which a fragment of a pseudogene in chimpanzee has been linked to an existing gene rather than a case of human specific gene duplication and subsequent subfunctionalization. Because of this we removed PtTAS2R5 from the final candidates list.

The three remaining candidates all showed a good alignment of each of their transcript products with the product of one of the subfunctionalization candidates.

5.3.4 Analysis of the final candidates

After this very strict elimination of ambiguous cases, we examined the three remaining candidates in detail. All of these candidates are single genes in the chimpanzee genome that are alternatively spliced and the different splice variants correspond to different human genes. We located the exon boundaries in the protein alignment and determined if the exon structure was conserved between the chimpanzee alternative forms and the human proteins they hit. We also checked the EST hits for each of them in order to determine if the splice sites annotated in the alternative chimpanzee forms were supported by ESTs (which indicate supporting evidence). Although the three genes were classified by EnsEMBL as KNOWN, there were no ESTs that hit these genes with an identity of 95% or more. This, however, is not surprising, as the number of chimpanzee ESTs

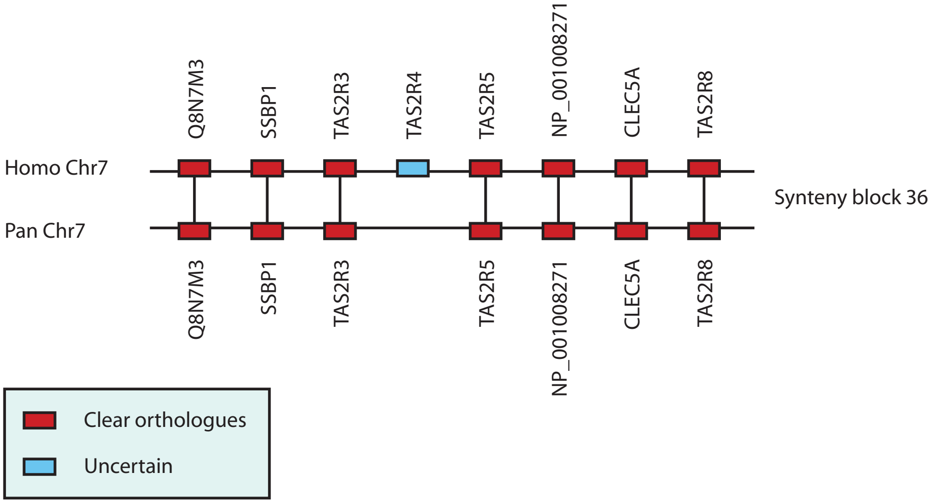


Figure 5.5: Genes surrounding the area where TAS2R5 is located in the human chimpanzee genome

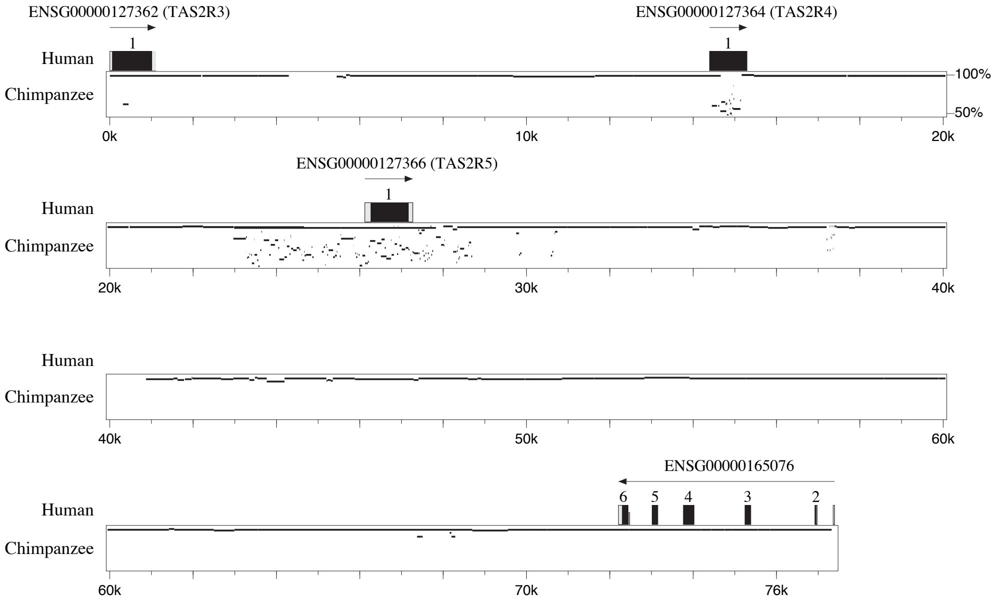


Figure 5.6: Genomic alignment of the the area where TAS2R5 is located in both human and chimpanzee

present in the database is only 6930.

SPAG11

The chimpanzee SPAG11 gene (PtSPAG11) is annotated as the Sperm-associated antigen 11 precursor and it is located on chimpanzee chromosome 8.

The human SPAG11 was initially characterized as the EP2 gene, and arose by the fusion of two ancestral β-defensin genes arranged in tandem. It is specifically expressed in the epididymis and it generates several different splice forms that may have little or no similarity to each other (Frohlich *et al.*, 2001). The SPAG11 group of genes are characterized by several alternatively spliced forms and some have been reported to use different reading frames in different transcripts, and indeed we observe this use of different reading frames in the transcripts that are different between the two human genes. These proteins are involved in the innate immune system and also in male reproductive functions (Yenugu *et al.* 2006, Hall *et al.* 2007).

According to the EnsEMBL annotation PtSPAG11 has six alternative transcripts, and is a one to many orthologue to the human genes SPAG11B (Sperm-associated antigen 11 precursor) and SPAG11A (Sperm associated antigen 11B isoform H precursor). Both are located on human chromosome 8 separated by roughly 200 kb. The two human genes have the exact same size (15,917 bp) and the sequence identity between them is 99.65%. Their identity with the chimpanzee gene which also has a similar size (15,942 bp) is 99.55% for SPAG11B and 98.36% for SPAG11A. This supports a very recent origin of these genes by a duplication on chromosome 8.

We examined the products of the six different transcripts that are encoded by PtSPAG11 by aligning them to those products from the two human SPAG11 genes. Three of the proteins predicted for PtSPAG11 are annotated for both of the human genes. The other three (figure 5.7) are annotated for only one of these two human genes. Two of them, ENSPTRP00000039008 and ENSP-

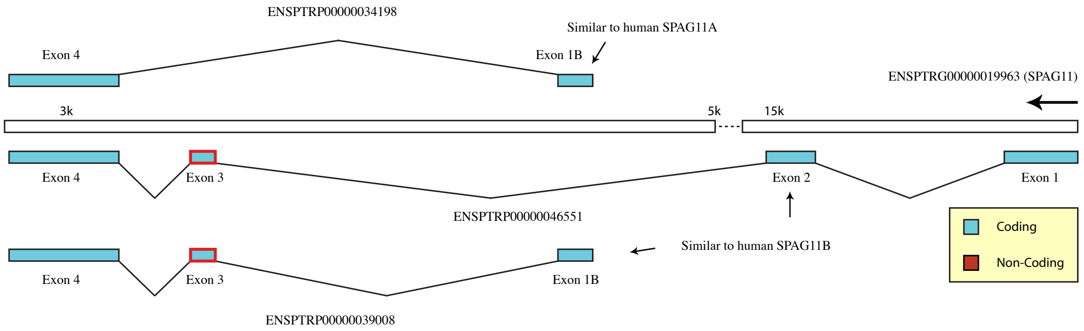


Figure 5.7: Schematic representation of the exon structure of the three chimpanzee alternative SPAG11 transcripts that correspond to the transcripts expressed only in one of two different human genes. Exon 3, marked in red, produces a frameshift when included in the transcript that causes a premature stop codon on exon 4

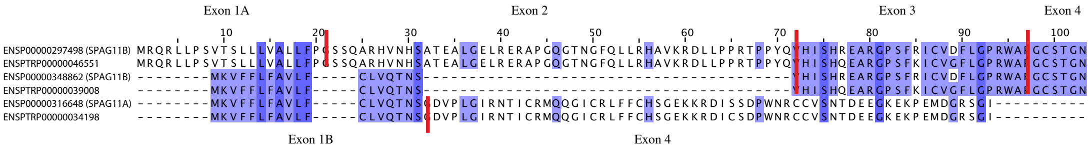

Exon 1A
Exon 2
Exon 3
Exon 4
10
20
30
40
50
60
70
80
90
100
ENSP00000297498 (SPAG11B) MRQRLLPSVTSLLLVALLFPGSSQARHVNHSATEALGELRERAPGQGTNGFQLLRHAVKRDLLPPRTPPYQYHISHREARGPSFRICVDFLGPRWAFGCSTGN
ENSPTRP00000046551 MRQRLLPSVTSLLLVALLFPGSSQARHVNHSATEALGELRERAPGQGTNGFQLLRHAVKRDLLPPRTPPYQYHISHQEARGPSFKICVGFLGPRWAFGCSTGN
ENSP00000348862 (SPAG11B) --------MKVFFLFAVLF-----CLVQTNS--YHISHREARGPSFRICVDFLGPRWAFGCSTGN
ENSPTRP00000039008 --------MKVFFLFAVLF-----CLVQTNS--YHISHQEARGPSFKICVGFLGPRWAFGCSTGN
ENSP00000316648 (SPAG11A) --------MKVFFLFAVLF-----CLVQTNSGDVPLGIRNTICRMQQGICRLFFCHSGEKKRDISSDPWNRCCVSNTDEEGKEKPEMDGRSGI----------
ENSPTRP00000034198 --------MKVFFLFAVLF-----CLVQTNSGDVPLGIRNTICRMQQGICRLFFCHSGEKKRDICSDPWNRCCVSNTDEEGKEKPEMDGRSGI----------
Exon 1B
Exon 4

Figure 5.8: Alignment of the three *P. troglodytes* SPAG11 proteins –which are encoded by only one chimpanzee gene– with the human proteins encoded by two different genes SPAG11A and SPAG11B. intron positions are marked by red lines. In those cases in which the intron-exon boundary is in the middle of a codon, the line is plotted over the affected amino acid. The coding region of exon 4 has different sizes and does not align between the two forms because of the use of different reading frames.

TRP00000046551 correspond to two alternative forms annotated only for the human SPAG11B and the other form, ENSPTRP00000034198 corresponds to a form annotated only for the human SPAG11A. The alignment of these three chimpanzee splice variants and the corresponding human proteins is shown in figure 5.8 on the preceding page.

We investigated the molecular basis for the loss of different splice forms in the two human paralogues by examining the nucleotide alignment of these genes. We observed several differences in the exon sequences that give rise to differences in the coding region between the two genes, but we found no obvious difference between them that could lead to the loss of one of the alternative forms, such as the introduction of a stop codon in one of the genes that may cause the termination of translation in one frame, or the absence of a splice site.

The largest number of substitutions was in exon 3 (as numbered in figure 5.7 on the facing page) of SPAG11A, which corresponds to exon 3 of ENSP00000297498 and exon 2 of the ENSP00000348862 transcripts. In the gene SPAG11B, which doesn't present this exon in any of its annotated forms, the sequence surrounding this exon was identical to the chimpanzee gene sequence, which is predicted to have the splice form. Because of the lack of EST data however we cannot confirm that the chimpanzee form is actually expressed.

The previous publications regarding SPAG11 mention only one human SPAG11 gene with many alternatively spliced variants, but do not mention the existence of another human SPAG11 gene. In order to find out if there is any logical explanation why this second gene may have been missed we examined it further. We examined the synteny blocks described in chapter 3 on page 55 for chromosome 8, however the area containing this gene was not within any of the human-chimpanzee synteny blocks in either organism. We aligned the human and chimpanzee regions of chromosome 8 that contained the SPAG11 genes in both organisms using MultiPipMaker (Schwartz *et al.*, 2000); Figure 5.9 on page 153. We confirmed the existence of several duplications in this area. One of them, an inverted segmental duplication specific to the human genome, contained the two copies of SPAG11 in this species (figure 5.9 overleaf). The presence of these duplications also explains the lack of a synteny block in our analysis in this area between the human and chimpanzee genomes, as there would be no unambiguous one-to-one synteny within the region.

We hypothesize that because the sequence of these two genes is highly similar, differences in the EST sequences from the two genes may have been mistaken for sequencing errors in the SPAG11 literature, particularly if the existence of a second gene was not suspected, and the location of both duplicates within 200 kb of each other may prevent their resolution as two separate genes with the methods used at the time it was characterized.

Both SPAG11 (Frohlich *et al.*, 2001) and DEFB2, another member of the defensin family (Harder *et al.*, 1997), are located within the same block that has been duplicated in human. The position of both of these genes was initially obtained using the same yeast artificial chromosome (YAC) mapping technique (and some of the same YACs). This technique involves the use of overlapping YACs that have certain sequence-tagged sites (STS) at known locations along them. These STS allow estimation of the location of the gene of interest by finding a set of YACs from which the gene may be amplified by polymerase chain reaction (PCR) and determining the minimum overlapping area between them. From the STS located in this overlapping area the location of the gene of interest can be deduced. The accuracy of this technique depends on the separation between the STS that are used along the YACs, which is the factor that will determine the maximum resolution of a location determined in this manner assuming there are no problems with the integrity of the YACs. The separation between the two copies of SPAG11 on the human chromosome 8 is roughly 200 kb. This number is roughly the same as the average STS resolution from the STS-map of the human genome, but much lower than the real effective resolution which is closer to 1 Mb (Lauer *et al.*, 1998). These two issues could

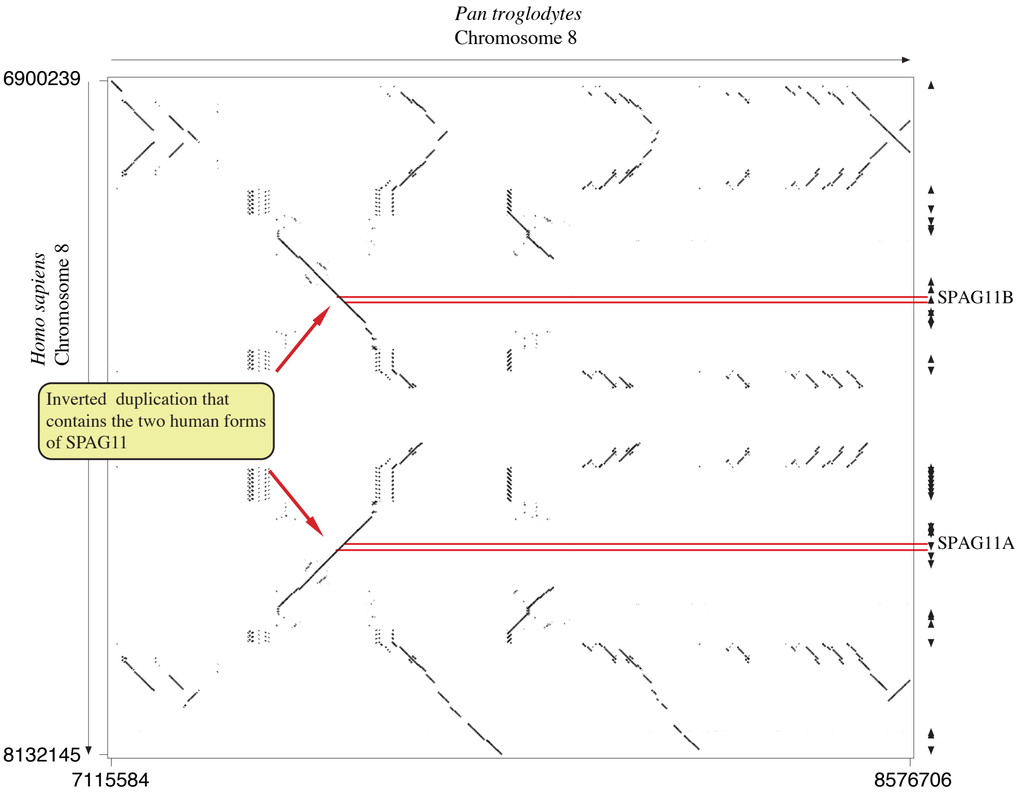


Figure 5.9: Dotplot representation of the alignment between the chimpanzee region of Chromosome 8 that contains PtSPAG11 and the region of chromosome 8 in human that contains the two recently duplicated versions of the human SPAG11. The numbers on the axes represent the length of the sequence and the arrows on the vertical axis indicate the position of the human genes, with the two SPAG11 genes indicated on the dotplot.

explain the lack of mention of the second copy of the gene in the literature, assuming the duplication is real and there is not a major assembly error in the human chromosome 8 which is unlikely given its high quality.

Our results suggest both human SPAG11 genes originated from a very recent recent segmental duplication that has occurred in the last 6 My. This is a clear case in which a gene that undergoes alternative splicing has been duplicated since the divergence of the human and chimpanzee lineages. However we find no obvious explanation for why both human genes would not form all of the different alternatively spliced variants. Regardless of this, because this gene is involved in the immune system, and possibly in reproduction, as it is present on the surface of the sperm cells (Hall *et al.* (2007)), this gene is a very interesting candidate for future studies, as both immunity and reproduction play a very important role in the speciation process.

PRKX and PRKY

The human PRKX and PRKY genes, are located on the X and Y chromosomes, respectively. The chimpanzee gene PRKY is located on the chimpanzee Y chromosome. This gene has two alternative forms annotated that differ at the C-terminal part of the protein (figure 5.10 overleaf). The ENSPTRP00000055043 form corresponds to the PRKX gene in human and ENSPTRP00000037059 corresponds to the PRKY gene in human (figure 5.14 on page 162). The first exon of the chimpanzee transcript that encodes ENSPTRP00000055043 appears to be truncated and missing the initial Met according to the translation obtained from the ENSEMBL database. This appears to be an error in the annotation of the initial amino acid of that protein and not in the exon boundary prediction, as the initial exon annotated for both of the forms is the same.

In the case of human PRKY the gene has two transcripts annotated, but these are identical, with the exception of a truncation in the non-coding area of one of the exons, so for our purposes we consider this as a single transcript.

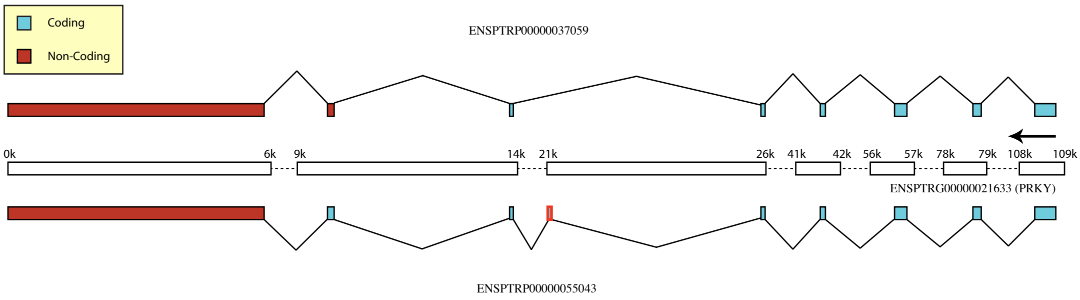


Figure 5.10: Schematic representation of the exon structure of the two candidate alternative transcripts belonging to the *P. troglodytes* PRKY gene. The critical difference between the two transcripts is highlighted with red box. The exclusion of this exon introduces a frameshift that results in the truncation of the protein at a stop codon in exon 7

The two alternative forms of the chimpanzee PRKY (PtPRKY) share the first 5 exons, and the difference between them is caused by the skipping of the 6th exon in one of the forms, this alters the reading frame and produces a premature stop codon 16 nucleotides downstream from the splice site. The protein alignment of the different forms can be seen in figure 5.11 overleaf.

The length of the genes is slightly different. The PRKY gene is slightly shorter, (107,576 bp) than the PRKX gene (109,246 bp) or the PtPRKY gene (108,825 bp).

We compared the genes using the MultiPipMaker tool (Schwartz *et al.*, 2000) and found a deletion in the PRKY gene that completely removed the exon 6 while this exon is still present in PRKX. As noted above, in the absence of this exon it is not possible to build the longer form, because a stop codon is introduced in the same reading frame, and even if exon 7 was completely skipped there is also a stop codon close to the start of the next exon in the same reading frame.

The similarity between the two human PRK genes which are 94% identical has previously been reported by Schiebel *et al.* (1997) who also reported the existence of two other pseudogenized copies, of these genes; one in the X chromosome and one in chromosome 15. In their work they describe the region spanning this

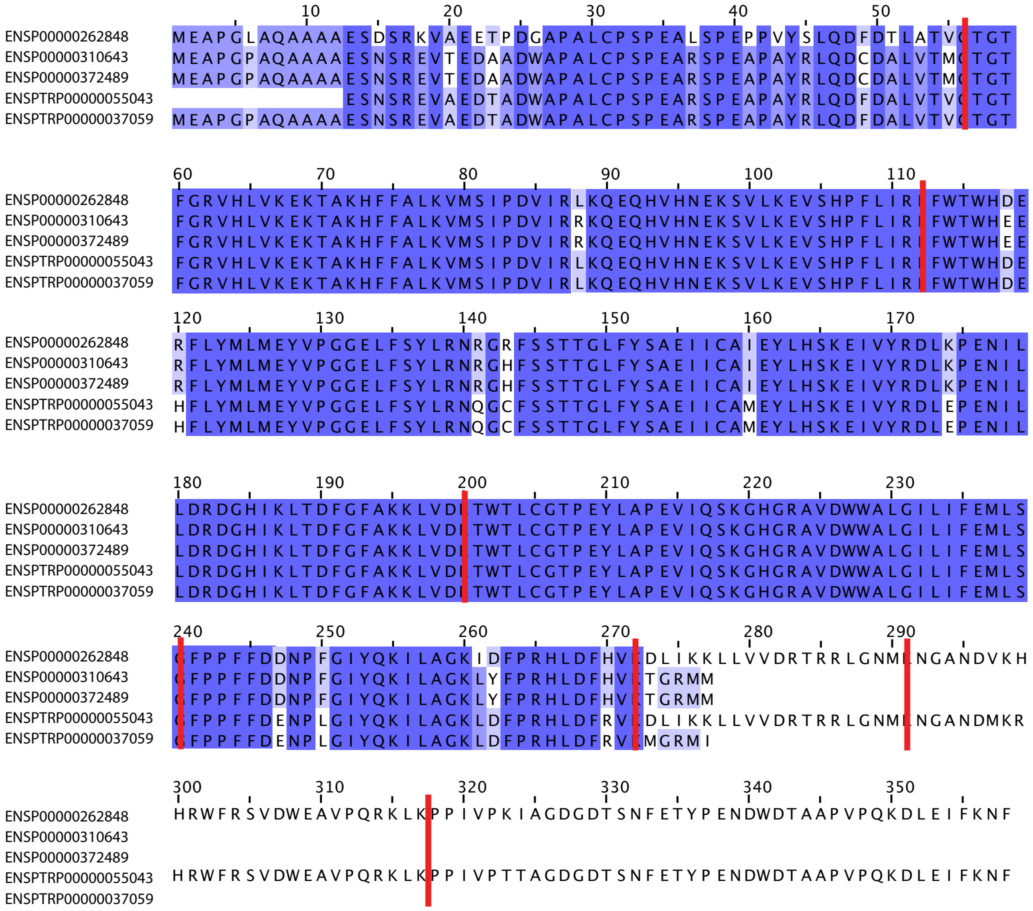

10 20 30 40 50
ENSP00000262848 MEAPGLAQAAAAESDSRKVAEETPDGAPALCPSPEALSPEPPVYSLQDFDTLATVCTGT
ENSP00000310643 MEAPGPAQAAAAESNSREVTEDAADWAPALCPSPEARSPEAPAYRLQDCDALVTMCTGT
ENSP00000372489 MEAPGPAQAAAAESNSREVTEDAADWAPALCPSPEARSPEAPAYRLQDCDALVTMCTGT
ENSPTRP00000055043 ESNSREVAEDTADWAPALCPSPEARSPEAPAYRLQDFDALVTVCTGT
ENSPTRP00000037059 MEAPGPAQAAAAESNSREVAEDTADWAPALCPSPEARSPEAPAYRLQDFDALVTVCTGT
60 70 80 90 100 110
ENSP00000262848 FGRVHLVKEKTAKHFFALKVMSIPDVIRLKQEQHVHNEKSVLKEVSHPFLIRIFWTWHDE
ENSP00000310643 FGRVHLVKEKTAKHFFALKVMSIPDVIRRKQEQHVHNEKSVLKEVSHPFLIRIFWTWHEE
ENSP00000372489 FGRVHLVKEKTAKHFFALKVMSIPDVIRRKQEQHVHNEKSVLKEVSHPFLIRIFWTWHEE
ENSPTRP00000055043 FGRVHLVKEKTAKHFFALKVMSIPDVIRLKQEQHVHNEKSVLKEVSHPFLIRIFWTWHDE
ENSPTRP00000037059 FGRVHLVKEKTAKHFFALKVMSIPDVIRLKQEQHVHNEKSVLKEVSHPFLIRIFWTWHDE
120 130 140 150 160 170
ENSP00000262848 RFLYMLMEYVPGGELFSYLRNRGRFSSTTGLFYSAEIICAIEYLHSKEIVYRDLKPENIL
ENSP00000310643 RFLYMLMEYVPGGELFSYLRNRGHFSSTTGLFYSAEIICAIEYLHSKEIVYRDLKPENIL
ENSP00000372489 RFLYMLMEYVPGGELFSYLRNRGHFSSTTGLFYSAEIICAIEYLHSKEIVYRDLKPENIL
ENSPTRP00000055043 HFLYMLMEYVPGGELFSYLRNQGCFSSTTGLFYSAEIICAMEYLHSKEIVYRDLEPENIL
ENSPTRP00000037059 HFLYMLMEYVPGGELFSYLRNQGCFSSTTGLFYSAEIICAMEYLHSKEIVYRDLEPENIL
180 190 200 210 220 230
ENSP00000262848 LDRDGHIKLTDFGFAKKLVDITWTLCGTPEYLAPEVIQSKGHGRAVDWWALGILIFEMLS
ENSP00000310643 LDRDGHIKLTDFGFAKKLVDITWTLCGTPEYLAPEVIQSKGHGRAVDWWALGILIFEMLS
ENSP00000372489 LDRDGHIKLTDFGFAKKLVDITWTLCGTPEYLAPEVIQSKGHGRAVDWWALGILIFEMLS
ENSPTRP00000055043 LDRDGHIKLTDFGFAKKLVDITWTLCGTPEYLAPEVIQSKGHGRAVDWWALGILIFEMLS
ENSPTRP00000037059 LDRDGHIKLTDFGFAKKLVDITWTLCGTPEYLAPEVIQSKGHGRAVDWWALGILIFEMLS
240 250 260 270 280 290
ENSP00000262848 GFPPFFDDNPFGIYQKILAGKIDFPRHLDFHVKDLIKKLLVVDRTRRLGNMINGANDVKH
ENSP00000310643 GFPPFFDDNPFGIYQKILAGKLYFPRHLDFHVKTGRMM
ENSP00000372489 GFPPFFDDNPFGIYQKILAGKLYFPRHLDFHVKTGRMM
ENSPTRP00000055043 GFPPFFDENPLGIYQKILAGKLDFPRHLDFRVKDLIKKLLVVDRTRRLGNMINGANDMKR
ENSPTRP00000037059 GFPPFFDENPLGIYQKILAGKLDFPRHLDFRVKMGRMI
300 310 320 330 340 350
ENSP00000262848 HRWFRSVDWEAVPQRKLKPPIVPKIAGDGDTSNFETYPENDWDTAAPVPQKDLEIFKNF
ENSP00000310643
ENSP00000372489
ENSPTRP00000055043 HRWFRSVDWEAVPQRKLKPPIVPTTAGDGDTSNFETYPENDWDTAAPVPQKDLEIFKNF
ENSPTRP00000037059

Figure 5.11: Alignment of the *P. troglodytes* PRKY gene products with the products of the two human genes PRKX (ENSP00000262848) and PRKY (ENSP00000310643 and ENSP00000372489).

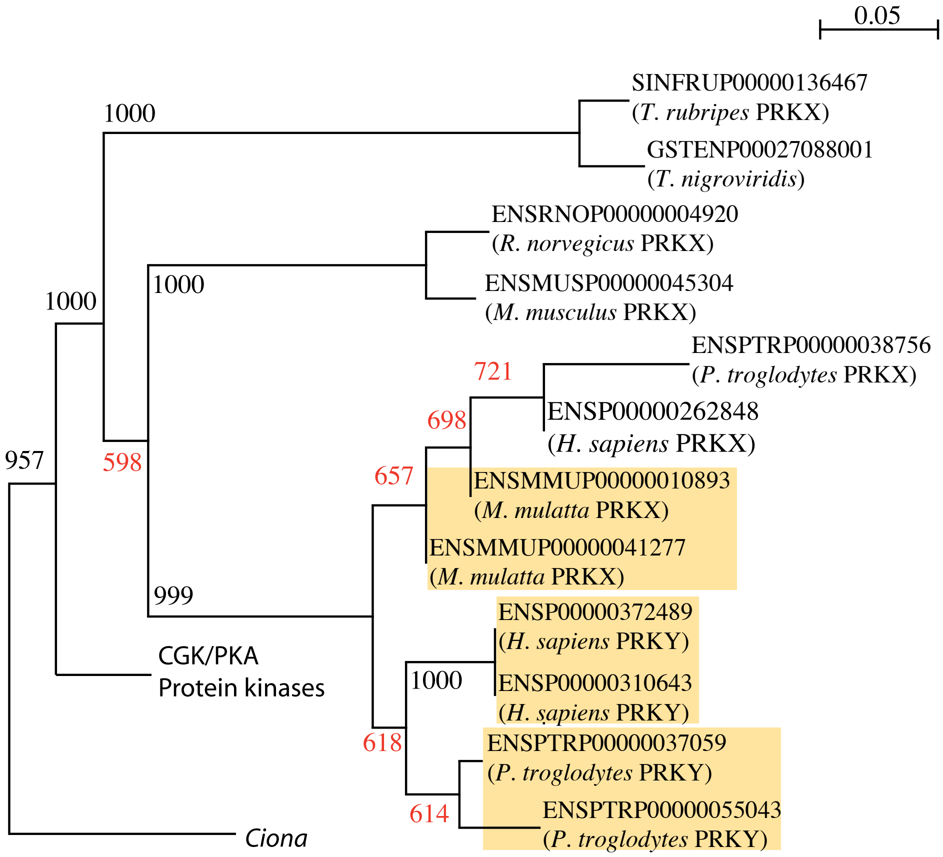


Figure 5.12: Gene tree including *P. troglodytes* PRKX and PRKY as well as the other members of the family. Those transcripts that are encoded by the same gene (MmPRKX, HsPRKY and PtPRKY) are indicated by a shaded box. The low bootstraps highlighted in red indicate that the branches are not reliable and the relationship is not completely resolved in some areas of the tree.

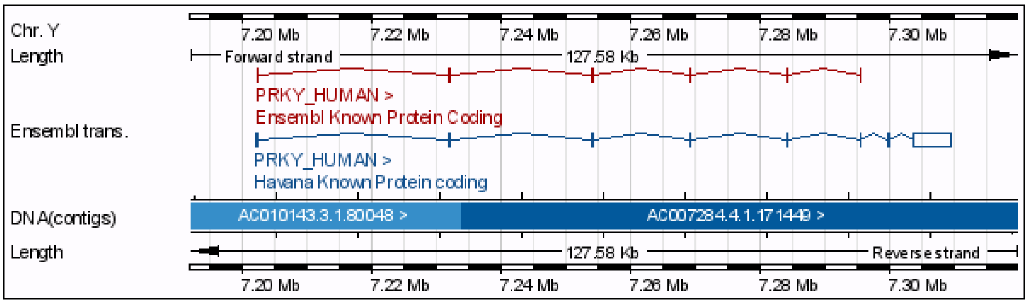


Figure 5.13: This figure shows the contigs in human that span the human PRKY gene. The area containing the PRKY exon deletion is contained completely within one of the human contigs. This picture was obtained from the EnsEMBL genome browser

gene as a hotspot for non-pseudoautosomal recombination between the X and Y chromosomes. This type of recombination leads to phenotypically recognizable features, namely XX male and XY female phenotypes in human. The PRKY-containing subregion exhibits the largest recombination rate of this type outside of the pseudoautosomal region of the Y chromosome, which is the only area that usually undergoes recombination during meiosis. The missing exon in the PRKY gene is also mentioned in this work, where it is inferred to be recent, as no further frameshifts or mutations have accumulated on the downstream exons, even though they can no longer be translated, and thus are likely free from purifying selection.

The human PRK-encoded proteins contain an ATP binding domain and a catalytic domain highly homologous to protein kinases. The short form does not include an arginine residue that plays a role in the assembly of the three-dimensional structure on known protein kinases (Schiebel *et al.*, 1997). When examined using the SMART database, the kinase catalytic domain in the Pt-PRKY is predicted to be inactive in chimpanzee because of the lack of one of the catalytic residues, a K (Lysine) at position 174 in the amino acid sequence. This is caused by the change of the nucleotide sequence from AAG (K) that is present in the human genes to GAG (Glutamic acid) that is present in the chimpanzee

PRKY.

We examined the EnsEMBL homology annotation and found the gene PtPRKY was annotated as one-to-one orthologue to the human PRKY and there was another chimpanzee gene, PtPRKX (ENSPTRG00000022476), annotated as a one-to-one orthologue to PRKX, which was located on the chimpanzee chromosome X. Both this gene and the human PRKX (ENSG00000183943) were located outside of the synteny blocks and because of this had not been identified as orthologues.

We examined this gene in order to determine why we had not identified it as the best hit of PRKX if it is indeed its orthologue. We found that the chimpanzee chromosome X region containing this gene has 19 sequencing gaps, seven of them larger than 1 kb in size. Most of them affect the intronic areas but one of them spans the whole exon 4, and the lack of this area makes the product of this gene fall outside of the RTH threshold in the BLAST search with the PtPRKY proteins. Because of the lack of the complete sequence of the gene we cannot be sure if this gene is functional or not in chimpanzees, although the regions for which the sequence is available are very highly conserved when compared with the human PRKX, and in this case the gene maintains the K residue that is predicted to be necessary for the kinase activity of the gene. However there is sufficient data to confirm, the PtPRKX gene does contain the exon 6 which is missing from its orthologue in human. This means the gene is capable of also producing the long ending of the protein, even if exon 4 was truly absent (and not just un-sequenced) because it would not disrupt the reading frame. So we can infer the exon was lost in the human PRKY form by a deletion that occurred within the last 6My.

In this case we conclude that the duplication occurred prior to the divergence of the human and chimpanzee lineage, and the genes are indeed evolving in a different manner, with PRKY in human losing the longer form and PRKY in chimpanzee retaining this longer form but possibly losing its kinase activity.

In order to determine when the duplication occurred we extracted from EnsEMBL the multiple sequence alignment of all the proteins that have been grouped into this family. This family, ENSF00000000732, described as cAMP dependent kinase catalytic subunit, consists of a group of cAMP and cGMP dependent protein kinases as well as the PRKX and PRKY genes. We removed from the set those incomplete sequences where the region available did not cover most of the alignment length or did not align well, this included the sequence of the macaque PRKY protein, of which only a fragment is available that is located in an unassembled scaffold, but not the chimpanzee PRKX which is missing exon four, as the rest of the sequence was complete.

We realigned the remaining sequences using t_coffee, and we used clustalw to build a neighbor-joining protein tree with the sequences using *Ciona* as an outgroup. This tree was built considering also positions with gaps, because most of the differences between the alternative forms lie on the terminal extensions and if we do not include these regions many areas of the tree are unresolved. The EnsEMBL gene trees were not used, because they are built using the longest transcript for each gene and thus excluded the alternative forms which are the focus of this study.

The inferred tree showed two groups of sequences, one where all the PKAs grouped and one where the PKRX and Y grouped together, although the presence of several low bootstrap values indicate some of these branches may not be reliable and should be interpreted with care (figure 5.12 on page 157). The genes present in fish and rodent show only one splicing variant which corresponds to the long form in the primates, and the tree topology indicates that the duplication event occurred after the rodent divergence and prior to the macaque speciation.

These results allow us to place the duplication of the gene sometime after the divergence of the rodent and primates. Between the divergence of primates and rodents and the divergence of human and macaque, the gene was duplicated. According to the EnsEMBL gene tree it occurred after the divergence of the lemur but before the macaque and human lineages separated, some time between 55 and 24 Mya. The alternatively spliced form may have appeared before or after this event, we cannot be sure as the sequence information in macaque is incomplete.

After the human and chimpanzee lineages diverged the human copy of PRKY lost the exon 6, and thus lost the ability to generate the long form. There is no clear reason why the PRKX gene should not be able to generate the short form. The chimpanzee PRKY on the other hand suffered an A ⇒ G mutation that caused the likely loss of the kinase activity of the proteins it encodes by causing a K ⇒ E amino acid substitution in one of the catalytic residues. From the available sequence of the chimpanzee PRKX gene we know it contains exon 6 and it does not have the A ⇒ G mutation that could render the kinase domain inactive.

Although the duplication and subfunctionalization of this gene has not occurred after the divergence of the human and chimpanzee lineages, we can see it has followed different evolutionary routes in each of these species, with both human and chimpanzee PRKY apparently losing or reducing their kinase catalytic ability by different routes.

ENSPTRG00000021919

The chimpanzee gene ENSPTRG00000021919 has two alternative splice forms (figure 5.14) and it is annotated by EnsEMBL as a one to one orthologue to the human SSX1 gene and as a paralogue of the human SSX6 gene, which means it is related to this gene through a duplication event that occurred prior to the speciation between the two lineages. In the protein sequence alignment of the products of these genes the similarity between each of the alternative forms in chimpanzee and each of the human forms can be seen (figure 5.15).

When examining the EnsEMBL family ENSF00000004534 to which these genes belong we found ENSPTRG00000021919 is most similar to the SSX1, SSX6 and SSX8 human genes. There is also a chimpanzee SSX8 gene annotated which

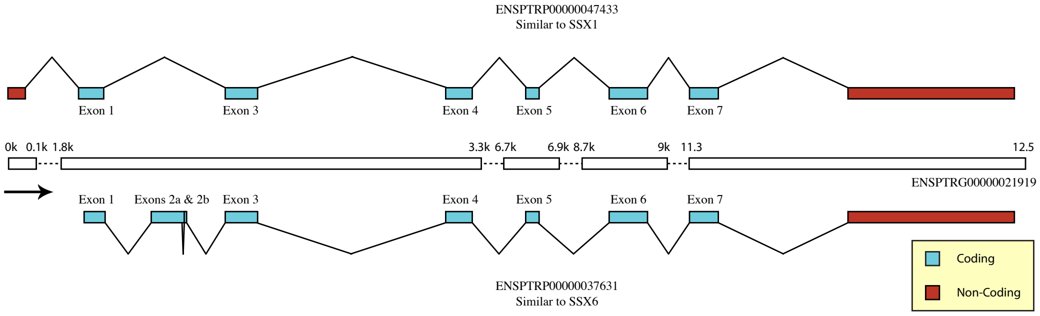
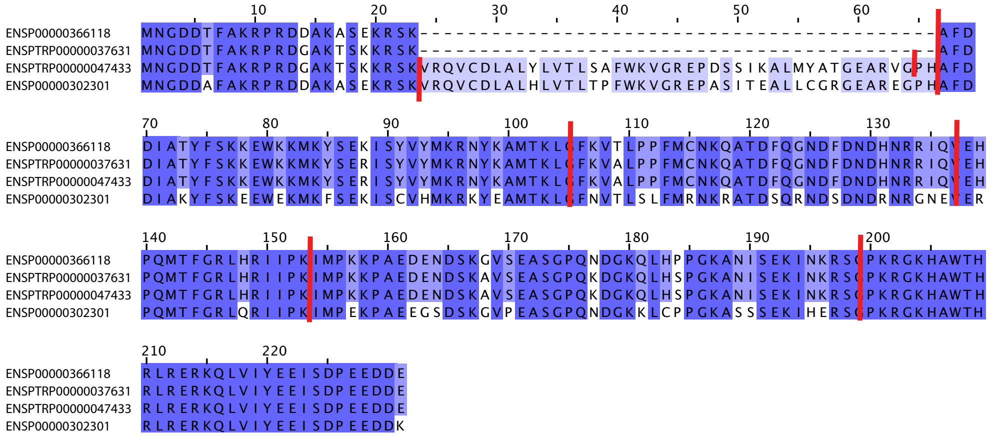


Figure 5.14: Schematic representation of the exon structure of the two candidate alternative transcripts belonging to the human ENSPTRG00000021919 gene. Exon numbers are indicated for the coding exons.


```
                   10        20        30        40        50        60
ENSP00000366118    MNGDDTFAKRPRDDAKASEKRSK-------------------------------------------AFD
ENSPTRP00000037631 MNGDDTFAKRPRDGAKTSKKRSK-------------------------------------------AFD
ENSPTRP00000047433 MNGDDTFAKRPRDGAKTSKKRSKVRQVCDLALYLVTLSAFWKVGREPDSSIKALMYATGEARVGPHAFD
ENSP00000302301    MNGDDAFAKRPRDDAKASEKRSKVRQVCDLALHLVTLTPFWKVGREPASITEALLCGRGEAREGPHAFD

                   70        80        90        100       110       120       130
ENSP00000366118    DIATYFSKKEWKKMKYSEKISYVYMKRNYKAMTKLGFKVTLPPFMCNKQATDFQGNDFDNDHNRRIQVEH
ENSPTRP00000037631 DIATYFSKKEWKKMKYSERISYVYMKRNYKAMTKLGFKVALPPFMCNKQATDFQGNDFDNDHNRRIQVEH
ENSPTRP00000047433 DIATYFSKKEWKKMKYSERISYVYMKRNYKAMTKLGFKVALPPFMCNKQATDFQGNDFDNDHNRRIQVEH
ENSP00000302301    DIAKYFSKEEWEKMKFSEKISCVHMKRKYEAMTKLGFNVTLSLFMRNKRATDSQRNDSDNDRNRGNEVER

                   140       150       160       170       180       190       200
ENSP00000366118    PQMTFGRLHRIIPKIMPKKPAEDENDSKGVSEASGPQNDGKQLHPPGKANISEKINKRSGPKRGKHAWTH
ENSPTRP00000037631 PQMTFGRLHRIIPKIMPKKPAEDENDSKAVSEASGPQKDGKQLHSPGKANISEKINKRSGPKRGKHAWTH
ENSPTRP00000047433 PQMTFGRLHRIIPKIMPKKPAEDENDSKAVSEASGPQKDGKQLHSPGKANISEKINKRSGPKRGKHAWTH
ENSP00000302301    PQMTFGRLQRIIPKIMPEKPAEEGSDSKGVPEASGPQNDGKKLCPPGKASSSEKIHERSGPKRGKHAWTH

                   210       220
ENSP00000366118    RLRERKQLVIYEEISDPEEDDE
ENSPTRP00000037631 RLRERKQLVIYEEISDPEEDDE
ENSPTRP00000047433 RLRERKQLVIYEEISDPEEDDE
ENSP00000302301    RLRERKQLVIYEEISDPEEDDK
```

Figure 5.15: Alignment of the *P. troglodytes* SSX gene products with the products of the two human genes SSX1 (top) and SSX6 (bottom).

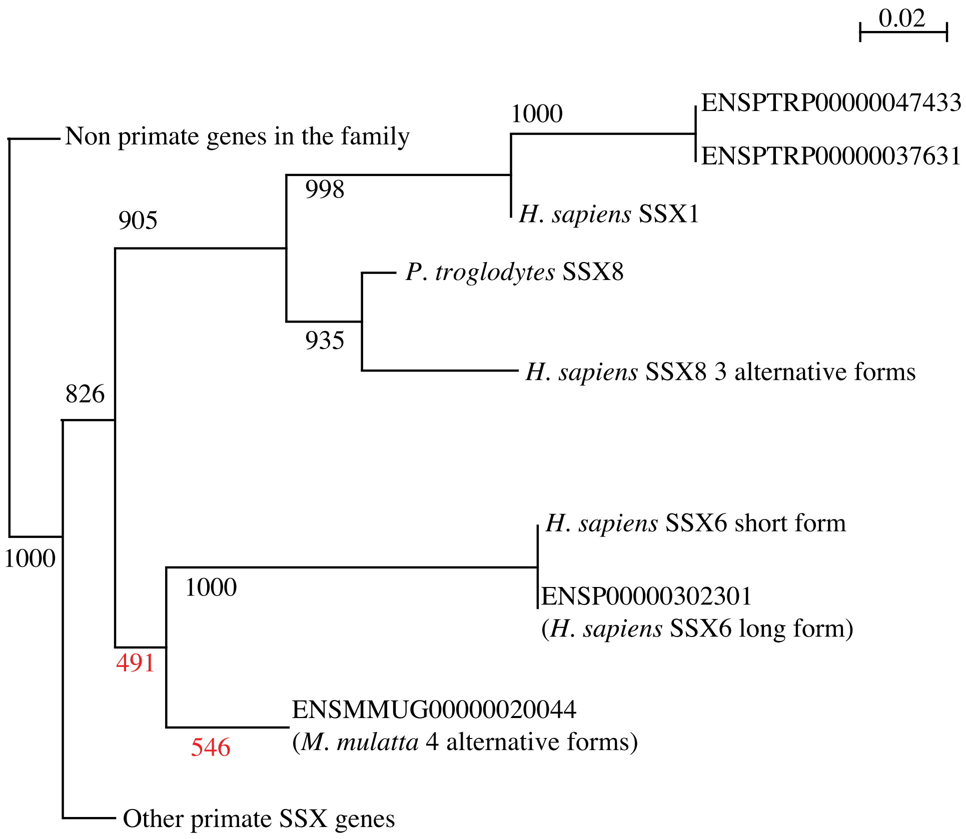


Figure 5.16: Gene tree including *P. troglodytes* ENSPTRG00000021919 as well as the closest genes from the SSX family. Some of the branches have a low bootstrap value, highlighted in red, and are not reliable.

only has one splice form. The human SSX6 gene has two splice forms; a short one and a long one, the same as this chimpanzee gene and the human SSX1 and SSX8 genes only have the short form.

We extracted from the EnsEMBL database the protein sequences for this family, and following the same procedure as with the PRKX family we removed those sequences that were incomplete or aligned badly. The remaining proteins were realigned and the resulting alignment was examined and corrected manually. Two trees were built, one without including the columns containing gaps and one including them. In both trees the topology of the subtree containing ENSPTRG00000021919 , PtSSX8, SSX1, SSX6 and SSX8 was the same, although the resolution of the alternative forms of each gene was better when including the areas with gaps. The resolution of the tree is quite bad overall, with low bootstrap values in many branches, however the resolution of the area that corresponds to these genes is better (figure 5.16). Although both alternative forms of ENSPTRG00000021919 group with human SSX1, this chimpanzee gene presents a long variant, which is only present in the SSX6 gene in human.

Most of the genes in this family belong to primates, only a few representatives from other species are present, with only one or two genes per species that in many cases align badly. There were no genes from Aves or earlier diverging lineages, which seems to indicate this family probably originated in mammals and recently underwent a large expansion within the primate lineage. In the three sequenced primates all the genes from this family are present in close clusters on chromosome X.

SSX genes in human have previously been described as a group that includes 9 genes and 10 pseudogenes, all of them located on two clusters on the X chromosome with the exception of one of the pseudogenes, ψSSX10 which is found on human chromosome 6 (Gure *et al.*, 2002). The large expansion of SSX genes in primates may have been caused by the duplication of a 100 kb region of the X chromosome that seems to have given birth to the second cluster, as there is only one mouse gene in the family according to the EnsEMBL classification. The cause of the premature stop in SSX8 responsible for the short form is an extra nucleotide in exon 7 that causes a frame shift resulting in the premature stop codon in this form (Gure *et al.*, 2002). When compared to the chimpanzee homologue we can see this is caused by 4 extra nucleotides in the human gene when compared to its chimpanzee homologue. Although EnsEMBL also provides two additional transcripts for SSX8 that do not have the premature stop codon, this is achieved by inserting a 4 bp intron, which is unlikely to be real.

When examining the reason why the long form of ENSPRTG00000021919 seems to have two exons instead of only the one exon 2 present in the human SSX6 long form we found there is an extra T inserted close to the end of the exon, and the annotation machinery inserts a highly unlikely 1 bp intron that restores the reading frame. This means that this candidate would only have one splice form.

SSX genes are normally expressed only in testis, but their expression has also been observed in different types of tumors. They have been involved in the t(X;18) translocation characteristically found in all synovial sarcomas Gure *et al.* (2002).

This case is still interesting, as we see how the chimpanzee has completely lost one of the splice forms, while the human retains it as an alternative form of SSX6. There is a long form annotated in macaque also, although it would require careful examination in order to determine if there are no misannotations.

5.4 Discussion

In this chapter we have developed an automated pipeline that can be used for the search of genes that are present in one organism as a single gene and in another as a pair of duplicated genes that have undergone a subfunctionalization process by which they have lost one of the alternatively spliced forms, but have retained between them the complete repertoire of ancestral alternative forms.

Our pipeline identified fourteen potential candidates out of the whole genomes of human and chimpanzee. These fourteen candidates were carefully examined and in ten cases the pipeline had identified a potential candidate due to the misannotation of what appeared to be more than one gene into a single gene in the database.

TAS25R was discarded as dubious because the gene belonged to a cluster of highly similar genes which could potentially lead to an erroneous assignment.

SSX was selected because of a misannotated transcript. Although here we see the loss of the long form of the gene product in chimpanzee and the conservation in both human and macaque, it is a case of lineage-specific loss of a splice form, not a case of alternative splicing with posterior subfunctionalization.

PRKY selection was caused by the lack of a completely finished chimpanzee sequence, because if the complete sequence of chimpanzee gene PRKX was in the database the human PRKX would not have hit PtPRKY as a TRH. Although this case is not a case of a duplication in the human lineage that gave rise to alternative forms by differential loss, it is still interesting. It shows a gene duplication that occurred after the rodent and primate divergence, and if the EnsEMBL data for the lemur is complete, after the divergence of the lineage leading to this species. We do see the differential subfunctionalization between human and chimpanzee. In the human genome this gene has lost one of the exons in one of its copies, whilst in both of the chimpanzee copies this exon remains. However an interesting observation is that in both the species the PRKY appears to have lost its kinase activity, in the human due to the loss of exon 6 and in the chimpanzee due to the mutation of one of the important catalytic residues in the active site. PRKX appears to be implicated in hematopoietic cell differentiation and also in morphogenesis. In mammals it is expressed in fetal kidneys but not adult kidneys, and its expression can affect development and migration of renal cells, the responses of different cell varies with the presence or absence of cAMP, and with the activity or inactivity of the kinase domain independently (Li *et al.*, 2002). This may indicate a certain advantage in increasing the copy number of the gene in non-kinase related functions it may perform, but also some deleterious effect caused by the excess kinase activity. However to confirm this would require a more detailed study of the gene in order to determine all the processes in which it is involved and what areas of the gene are important for each of them. An interesting observation is that in the human form of autosomal dominant polycystic kidney disease (ADPKD) the expression of PRKX persists after the fetal stage (Li *et al.*, 2002), this is one of the most common inherited diseases in human, and the duplication of this gene in the ancestor of the Catarrhini might have contributed to this with the increase of the number of gene copies which may have a dosage effect.

The case of SPAG11 seems to be a real case in which a gene that encodes many alternative forms has been duplicated very recently in the human lineage, within the last 6My. The annotation indicates non-shared transcripts. However, we failed to find a clear difference at the sequence level that may cause any of the two different human forms to lose any of the alternative forms encoded by the chimpanzee gene. This may be caused by the small amount of time elapsed since the duplication. Because SPAG11 is involved both in immunity and reproduction, as is suggested by the location of its product as well as that of several other defensins on the sperm surface, regardless of the possibility of subfunctionalization, as mentioned above, this gene could be an interesting candidate for potential species barrier formation mechanisms.

The fact that most of the genes identified by our pipeline that fit within the criteria for alternatively spliced genes which have suffered a posterior subfunctionalization appear to be caused by annotation errors is another indication for the need for a finished version of the chimpanzee genome in order to accurately identify the differences that have accumulated since the divergence of the human and chimpanzee lineages. Without access to a finished version of this genome many attempts to discover differences will obtain erroneous results caused by problems in the assembly or annotation, and in other cases ambiguous results, that are potential differences, but of which we cannot yet be sure. And any automated methods that attempt to predict these differences without any human curation are likely to identify many false positives.

Although the quality of the annotation requires much improvement, and a fully finished chimpanzee sequence is a necessity if we intend to discover the full range of differences between the two species, even with the problems arising from the quality of the data we managed to identify one good candidate, and with better data quality the strictness of the filters may be relaxed in order to identify more potential candidates.

We also see how fast changes in the splice structure of orthologous genes in the recently diverged human and chimpanzee lineages can introduce large changes in the proteins that are expressed by both organisms which is the case in the PRKX/Y and SSX genes. In this way small changes at the genome level cause a rapid divergence of the proteomes.

Gene duplication is one of the main forces that drive the evolution of the genome, subfunctionalization as described by Force *et al.* (1999) is more likely in more complex genes. As an alternatively spliced gene encodes more than one protein product it is likely to be involved in more than one function, and so we would expect it to be more complex than a gene with no alternative forms on average. The fact observed by Su *et al.* (2006) that duplicated genes and genes belonging to large families have fewer alternative forms than those genes that are single copies in the genome supports this. The duplication of the original alternatively spliced gene would allow for the functions of the gene to be divided between the duplicates. Because the gene is alternatively spliced, this already offers a division of the functions, that allows simple changes in large areas of the gene to completely remove one of the forms. For example, a single insertion in an alternative exon can alter the reading frame and render that form completely inactive, and because the area where this mutation can occur is quite large we would expect alternatively spliced genes to undergo subfunctionalization at a faster rate than those genes that do not have multiple forms, as the number of permissible mutations that would inactivate one of the functions without affecting the other are much larger.

Although two of the cases we examined (PRKX and SSX) were not duplications which had occurred after the speciation with chimpanzee, they are still relatively recent in evolutionary scale, and we can see clear differences between the duplicated copies. However in order to determine if there is a real difference between the rate of subfunctionalization of genes with alternative splice forms as would be expected if this idea is correct we will need to find many more examples, of both types, and for this kind of search a greater genome coverage than that provided by a draft sequence is fundamental in order to prevent incorrect results due to annotation errors or incomplete sequences.

Chapter 6

Rate of intron gain and loss in simultaneously duplicated *Arabidopsis thaliana* genes[1]

6.1 Introduction

The origin and evolution of introns in eukaryotic genomes has been hotly debated for many years. Central to these arguments is the question of how abundant intron gains and losses are. The evolution of introns is influenced by both mutation bias and selection. Intron length and intron number often appear to be affected independently. Mutation biases may cause positional biases of introns within a gene (Mourier and Jeffares 2003 proposed greater intron loss from the 3' ends of genes in intron-poor genomes, however, Nielsen *et al.* 2004 found conflicting results) and within a genome (long introns are rare in G+C rich regions; Duret *et al.* 1995). The selective effects of introns may be positive (facilitation of exon shuffling; Fedorov *et al.* 2003) or negative (additional transcriptional cost; Jeffares *et al.* 2006). Recently, it was shown that introns in *Arabidopsis thaliana* are shortened by selection for transcriptional efficiency (Seoighe *et al.*, 2005) mir-

[1] Knowles and McLysaght (2006)

roring a result found in other genomes (Castillo-Davis *et al.*, 2002). However, Lynch has argued that evolution of gene structure elements such as introns can be explained by neutral or nearly neutral evolution (Lynch and Richardson 2002, Lynch 2006).

Most previous studies of intron gain and loss have focused on identifying the prototypic gene structure in early eukaryotes and have thus examined this phenomenon in very distantly related eukaryotic genomes (some recent examples of large-scale studies include: Rogozin *et al.* 2003, Qiu *et al.* 2004, Rogozin *et al.* 2005, Roy and Gilbert 2005a). These broadly similar studies have returned strikingly different conclusions from an intron-rich ancestor with a preponderance of intron loss (Roy and Gilbert, 2005a) to a less intron-dense animal-plant ancestor, with gains outnumbering losses (Rogozin *et al.*, 2003). The different outcomes are probably due to differing assumptions about the properties of intron gain sites (Nguyen *et al.* 2005, Rogozin *et al.* 2005) or to different patterns of evolution in different lineages (Nielsen *et al.* 2004, Roy and Gilbert 2006). Indeed, a reanalysis of the Rogozin *et al.* (2003) data by Roy and Gilbert (2005a) using maximum likelihood methods instead of parsimony concluded that intron loss, and not gain, had dominated their evolution. Recently, Roy and Gilbert (2005b) estimated the rate of intron loss and gain to be 2×10^{-3} to 2×10^{-4} per million years and 6×10^{-7} to 4×10^{-6} per site per million years, respectively, based on comparisons across diverse eukaryotic lineages.

Studies of intron gain and loss in more recently diverged genomes include mammals (Roy *et al.*, 2003), *Caenorhabditis* (Coghlan and Wolfe, 2004), and fungi (Nielsen *et al.*, 2004). In a comparison of human and rodent introns, Roy *et al.* (2003) uncovered only loss events. The Coghlan and Wolfe (2004) study searched only for gain events and found evidence for 122 newly inserted introns that originated in the 80 - 110 Myr that separate *C. elegans* and *C. briggsae.* Nielsen *et al.* (2004) examined the patterns of intron evolution in fungi and uncovered a combination of intron loss and gain events.

The genome of the model plant *Arabidopsis thaliana* provides an ideal data set for examining intron gain and loss. Mounting evidence supports the occurrence of at least one, and likely multiple, whole-genome duplication events in the *Arabidopsis* lineage (AGI 2000, Blanc *et al.* 2000, Paterson *et al.* 2000, Vision *et al.* 2000, Simillion *et al.* 2002, Blanc *et al.* 2003, Bowers *et al.* 2003, Blanc and Wolfe 2004). The most recent of these genome duplication events is the most unequivocal, having generated a set of large blocks of duplicated genes that cover almost the entire genome with no overlap between blocks (Blanc *et al.* 2003; Bowers *et al.* 2003). Subsequent to the whole-genome duplication, many duplicated genes were lost and only approximately 2000 genes remain in duplicate today. The genes retained in duplicate are not a random sample of all genes and are biased for genes with a function in transcriptional regulation (Seoighe and Gehring, 2004).

Here we analyze the set of paralogous pairs of genes generated by this recent genome duplication for evidence of intron gain and loss in the period since the duplication event. These genes were all duplicated simultaneously and by the same mechanism. At the time of duplication, both paralogues had identical gene structures. This is not necessarily the case for paralogues that have duplicated by other means, for example, retrocopied genes (which are generated by the reverse transcription of mRNA and insertion of the cDNA into the genome) are usually completely devoid of introns at the time of duplication. We estimate the rates of intron loss and gain and test for a relationship with other properties of the genes concerned, such as expression level, G+C content, intragenic location, and function.

6.2 Materials and Methods

6.2.1 Duplicated *A. thaliana* genes

The sequences of the set of genes duplicated in the most recent whole-genome duplication as described by (Blanc *et al.*, 2003) were obtained from GenBank. One gene, At1g52000, was present in more than one duplicated pair and was excluded from further study. Sequences currently annotated as pseudogenes were also excluded. In nine cases, the locus ID had changed since the (Blanc *et al.*, 2003) study, and we replaced the old locus ID with that of the gene with identical sequence and location (determined by shared adjacency with at least one gene). The full list of gene pairs is available in Supplementary Table 1 at `http://mbe.oxfordjournals.org/cgi/content/full/msl017/DC1`.

6.2.2 Identification of non-conserved introns

A total of 2563 *A. thaliana* paralogues generated by a recent whole-genome duplication were aligned at the protein level using T-Coffee version 1.32 with default parameters (Notredame *et al.*, 2000). For each of the pairs, we identified the positions in the alignment corresponding to the intron splice site locations of each of the 23,164 introns in these genes. The quality of the alignment around the intron splice site was evaluated by examining ten alignment positions on each side of the splice site following the method used by (Coghlan and Wolfe, 2004). An unambiguous alignment region was defined as one with at least five conserved amino acids and no alignment gaps in the ten alignment positions on each side of the splice site (20 positions in total). An intron was classified as conserved if the location and phase were identical in the alignment of the two paralogues and if there were no other introns within 5 amino acids of this position on either side. An intron was classified as non-conserved if there was no intron in the paralogue in an identical position or within 5 amino acids in the alignment. Cases where the alignment was ambiguous, intron location but not phase was conserved, or where there was another intron within 5 amino acids on either side of the splice site were marked as ambiguous and excluded from further analysis.

6.2.3 Detection and alignment of plant homologous sequences

All *A. thaliana* sequence pairs with at least one non-conserved intron were used as queries in a BLAST search against genomic DNA of the *Viridiplantae* division of GenBank. The database was searched using tBLASTn with an expectation (e) value threshold of $1e \times 10^{-4}$ and only retaining hits with an e value within a range of $1e \times 10^{5}$ from the top non-*A. thaliana* hit.

Some of the retrieved hits were very long (e.g., entire chromosomes) which may feasibly contain more than one genuine homolog. For each hit, all of the high scoring pairs with e-values below the threshold were selected. The BLAST search also returned many sequence fragments that did not align with the whole *A. thaliana* gene or with the region surrounding the intron. These short fragments negatively affect the quality of the sequence alignment produced by automated methods. We implemented an iterative protocol to remove poorly aligned sequences and sequences that did not span the area of interest (i.e., the region of the intron) as follows:

Retrieved similar sequences were initially aligned to the already aligned *A. thaliana* pair using T-Coffee. In the first iteration, retrieved sequences that did not have at least five aligned bases in 30 bp on either side of the intron splice site were removed, and the remaining sequences were realigned. The resulting alignment was reexamined, sequences with fewer than ten aligned bases within 30 bp of the intron were removed and the sequences were realigned. In the third iteration, only sequences with 15 aligned bases within 30 bp of the intron position were retained. In each case, the 30-bp window is offset by 10 bp on each side to avoid the immediate region of the splice site that has a tendency to align poorly if there is an intron in one of the sequences (i.e., the splice site region aligns completely to one side or the other and not partially on each side of the intron as would be ideal). A final alignment was produced for each of the non-conserved *A. thaliana* introns and remaining homologous sequences.

6.2.4 Identification of homologous introns

An homologous intron was identified from an alignment as a stretch of at least 40 bp aligned between the -6 and $+6$ *A. thaliana* intron splice site nucleotides and aligned with gaps in the *A. thaliana* gene lacking the intron, that is, requiring that the intron is at least 28 bp long. We required that 10 base pairs on both sides of the splice site region (from -15 to -5 and from $+5$ to $+15$) should be aligned without gaps in order to unambiguously declare the presence or absence of a homologous intron. If one of the aligned sequences had gaps in this region, it was removed from the alignment. These cleaned alignments were used to construct a neighbor-joining tree for each non-conserved intron with ClustalW (Thompson *et al.*, 1994) using Kimuras correction for multiple hits and ignoring positions with gaps.

6.2.5 Similarity of introns and other regions of the *A. thaliana* genome

We used BLAST to search with the sequence of all non-conserved introns against the genome of *A. thaliana* without filtering low complexity regions and with an e value threshold of 1. In order to recover any hits that might be missed by the BLAST method, we also used SSearch with the threshold set to 0.1 and default parameters (Pearson, 1996). We discarded the self-hits and those hits with a length of less than 50% of the query sequence, this removed most of the hits due to repeats in the sequence. In order to remove those hits that were due to a large-scale duplication (whole-gene duplication or segmental genome duplication), we removed hits where the similarity extended for long regions outside the intron sequence.

6.2.6 Difference in expression levels between genes with gained and lost introns

Affymetrix data from 11 microarrays corresponding to expression levels in leaf (3), stem (4), and flower (4) for growth in two different conditions – greenhouse and growth chamber – were downloaded from the GEO Website (`http://www.ncbi.nlm.nih.gov/projects/geo`). All the genes for which we had expression data were classified into 10 equal-sized expression categories. The data from the same tissue in the equal growth conditions were pooled before analysis. Using only those genes in which all introns had been classified as gained, lost, or conserved, we examined if those genes with gained or lost introns were more abundant in certain expression categories using a χ^2 test.

6.2.7 Distribution of gains and losses along the coding sequence

Intergene method

All introns were classified into 10 different location categories according to their relative position along the coding sequence (CDS) of the gene. Category 1 indicates that the intron was in a position between 0% and 10% of the length of the gene, category 2 indicates that the intron was in a position 11 - 20% along the gene, etc. We examined if there was any significant difference in the distribution along the coding sequence between gained, lost, and conserved introns using a χ^2 test.

Intragene method

The intragenic location of intron gains and losses was also examined on a per gene basis as per the method of Lin and Zhang (2005). Each gene was classified as one of unbiased, 5' biased, or 3' biased based on the relative number of introns in the 5' or the 3' half of the gene. The null expectation is that the number of genes with a 5' bias should equal the number with a 3' bias, and this was tested using a chi-squared test. This was done separately for all non-conserved introns (in 486 genes), for gained introns, and for lost introns.

6.2.8 Examination of functional bias in gene with non-conserved introns

GOslim annotation data for the genes in the *A. thaliana* genome were downloaded from The Arabidopsis Information Resource (TAIR) Website (Berardini *et al.*, 2004) on 10 December 2005. Each gene pair was assigned the combined GOslim terms of each of its genes. For the purposes of this analysis, we excluded 636 gene pairs that contained no non-conserved introns and at least one ambiguous intron because we cannot be sure whether these are cases of conserved or non-conserved gene structure. This resulted in a set of 1927 gene pairs which we could definitively say did or did not experience an intron insertion/deletion (indel).

The expected frequencies of GOslim terms among the 281 gene pairs with at least one non-conserved intron were determined using simulations. We randomly sampled 281 gene pairs from the 1927 paralogous genes in our data set and noted the distribution of GOslim terms. This was repeated 100,000 times. The mean and standard deviation (SD) of the frequency of each GOslim term was calculated for the simulations and compared with the observed data. This procedure was repeated for the gene pairs with at least 2 and with at least 3 non-conserved introns. Correction for multiple tests was done in two alternative ways: Bonferroni correction and Benjamini-Hochberg correction (Benjamini and Hochberg, 1995).

We performed another set of 100,000 simulations for genes with at least 1, 2, and 3 non-conserved introns correcting for number of introns in the gene pair. In the simulations, for each of the original 281 gene pairs, a gene pair was selected randomly from the list of pairs with at least one member of the selected pair having the same number of introns as one member of the original pair.

6.2.9 Comparison of TCH3 with CAM3

The dotplots used in order to compare the sequences of these genes were generated using EMBOSS dotmatcher with a window size of 40 and a threshold score of 60.

6.3 Results

6.3.1 Recent Changes in *A. thaliana* Gene Structure

We examined 2563 paralogous *A. thaliana* gene pairs as identified by Blanc *et al.* (2003) originating from the recent whole-genome duplication 2060 MYA for changes in the presence or absence of introns. We aligned the paralogues using T-Coffee (Notredame *et al.*, 2000) and compared the alignment locations of introns within each pair. We employed stringent criteria to evaluate the quality of the alignment as per Coghlan and Wolfe (2004), and only introns in unambiguous portions of the alignment were considered further (see Methods 6.2 on page 174).

Conserved introns were defined as those present at an identical alignment location in each paralogue. Non-conserved introns are those with no intron in the corresponding location in the paralogue or within a short distance of that location.

We identified 10,004 pairs of introns that have been conserved in both *A. thaliana* paralogues since the genome duplication, 578 non-conserved introns (Supplementary Table 2 at `http://mbe.oxfordjournals.org/cgi/content/full/msl017/DC1`), and 2578 ambiguous cases. The 578 non-conserved introns are the results of either intron gain into one paralogue or loss from the other. We found 281 genes having one non-conserved intron each. An additional 115 gene

Table 6.1: Gene pairs with three or more non-conserved introns. Introns are classified as Non-conserved (N), Conserved (C) or Ambiguous (A)

Number of introns			Gene A	Description	Gene B	Description
N	C	A				
9	0	3	At3g09900	Ras-related GTP-binding protein, putative	At5g03530	Ras family GTP-binding protein
7	10	1	At2g21520	SEC14 cytosolic factor, putative / phosphoglyceride transfer protein, putative	At4g39170	SEC14 cytosolic factor, putative / phosphoglyceride transfer protein, putative
6	0	1	At3g48750	cell division control protein 2 homolog A (CDC2A)	At5g63610	protein kinase, putative
6	0	1	At1g15080	phosphatidic acid phosphatase family protein / PAP2 family protein	At2g01180	phosphatidic acid phosphatase family protein / PAP2 family protein
5	6	2	At4g28220	NADH dehydrogenase-related	At2g20800	pyridine nucleotide-disulphide oxidoreductase family protein
5	38	0	At1g80490	WD-40 repeat family protein	At1g15750	WD-40 repeat family protein
5	0	9	At4g17890	human Rev interacting-like family protein / hRIP family protein	At5g46740	ubiquitin-specific protease 21 (UBP21)
5	26	0	At4g02570	cullin family protein	At1g02980	cullin family protein
5	0	5	At1g76360	protein kinase, putative	At1g20650	protein kinase family protein
4	12	1	At1g30810	transcription factor jumonji (jmj) family protein / zinc finger (C5HC2 type) family protein	At2g34880	transcription factor jumonji (jmj) family protein / zinc finger (C5HC2 type) family protein
4	0	4	At3g55600	expressed protein	At2g39790	mitochondrial glycoprotein family protein / MAM33 family protein
4	0	6	At1g05900	endonuclease-related	At2g31480	expressed protein
4	32	0	At1g30820	CTP synthase, putative / UTP–ammonia ligase, putative	At2g34890	CTP synthase, putative / UTP–ammonia ligase, putative
3	2	6	At4g38550	expressed protein	At2g20960	expressed protein
3	4	10	At5g46380	hypothetical protein	At4g18150	hypothetical protein
3	24	3	At1g55970	histone acetyltransferase 4 (HAC4)	At3g12980	histone acetyltransferase 5 (HAC5)

Continued on next page

Table 6.1 – continued from previous page

Number of introns			Gene A	Description	Gene B	Description
N	C	A				
3	0	2	At1g09080	luminal binding protein 3 (BiP-3) (BP3)	At2g32120	heat shock protein 70 family protein / HSP70 family protein
3	4	0	At5g53400	expressed protein	At4g27890	nuclear movement family protein
3	8	0	At2g22660	glycine-rich protein	At4g37900	glycine-rich protein
3	8	2	At4g12030	bile acid:sodium symporter family protein	At4g22840	bile acid:sodium symporter family protein
3	2	0	At3g05960	sugar transporter, putative	At5g26340	hexose transporter, putative
3	6	6	At5g22650	histone deacetylase-related protein	At3g44750	histone deacetylase, putative (HD2A)
3	0	1	At5g66230	expressed protein	At3g51230	hypothetical protein
3	2	3	At5g06150	cyclin (cyc1b)	At3g11520	cyclin, putative (CYC2)
3	12	0	At5g40640	expressed protein	At3g27390	expressed protein
3	24	10	At1g73860	kinesin motor protein-related	At1g18410	kinesin motor protein-related
3	18	0	At4g26270	phosphofructokinase family protein	At5g56630	pyrophosphate-dependent phosphofructo-1-kinase-related protein
3	12	0	At4g12430	trehalose-6-phosphate phosphatase, putative	At4g22590	trehalose-6-phosphate phosphatase, putative
3	0	2	At1g74950	expressed protein	At1g19180	expressed protein
3	14	2	At1g11950	transcription factor jumonji (jmjC) domain-containing protein	At1g62310	transcription factor jumonji (jmjC) domain-containing protein
3	4	3	At1g01010	no apical meristem (NAM) family protein	At4g01550	no apical meristem (NAM) family protein
3	20	0	At5g27540	GTP-binding protein - related	At3g05310	GTP-binding protein-related
3	8	0	At3g09840	cell division cycle protein 48 (CDC48A) (CDC48)	At5g03340	transitional endoplasmic reticulum ATPase -related
3	20	1	At1g18870	isochorismate synthase, putative / isochorismate mutase, putative	At1g74710	isochorismate synthase 1 (ICS1) / isochorismate mutase
3	0	1	At4g26540	protein kinase family protein	At5g56040	leucine rich repeat protein kinase, putative
3	6	0	At1g70710	endo-1,4-beta-glucanase (EGASE) / cellulase	At1g23210	glycosyl hydrolase family 9 protein

Continued on next page

Table 6.1 – continued from previous page

Number of introns			Gene A	Description	Gene B	Description
N	**C**	**A**				
3	10	3	At2g18730	diacylglycerol kinase, putative	At4g30340	diacylglycerol kinase family protein

pairs have experienced multiple intron insertion/deletions (indels) in the time since duplication, 37 of which had 3 or more intron gains or losses (table 6.1).

6.3.2 Identification of Intron Gains and Losses

To distinguish intron gains from losses, we required genomic sequence data from homologous plant genes. We searched the *Viridiplantae* division of GenBank for similar flowering plant genomic DNA sequences spanning the intron position and aligned them to the pair of *A. thaliana* genes (section 6.2 on page 174). We again employed very stringent criteria on the quality of the sequence alignments. The most important criterion was the exclusion of alignments where there were gaps close to the intron position. These gaps may be indicative of poor alignment quality, thus making it impossible to confidently discern the presence or absence of an intron at the site of interest. An intron was inferred to have originated in the common ancestor of all genes containing the intron.

An intron gain was scored when the non-conserved intron was present only in one *A. thaliana* genome-duplication paralogue and other paralogues of this gene that duplicated after the tetraploidy event (figure 6.1 overleaf).

Evidence for intron loss comes from the presence of an intron in the same location in any earlier diverging flowering plant gene (figure 6.1).

We could confidently assign 56 intron gain events and 39 intron loss events (Supplementary Table 2 at `http://mbe.oxfordjournals.org/cgi/content/full/msl017/DC1`).

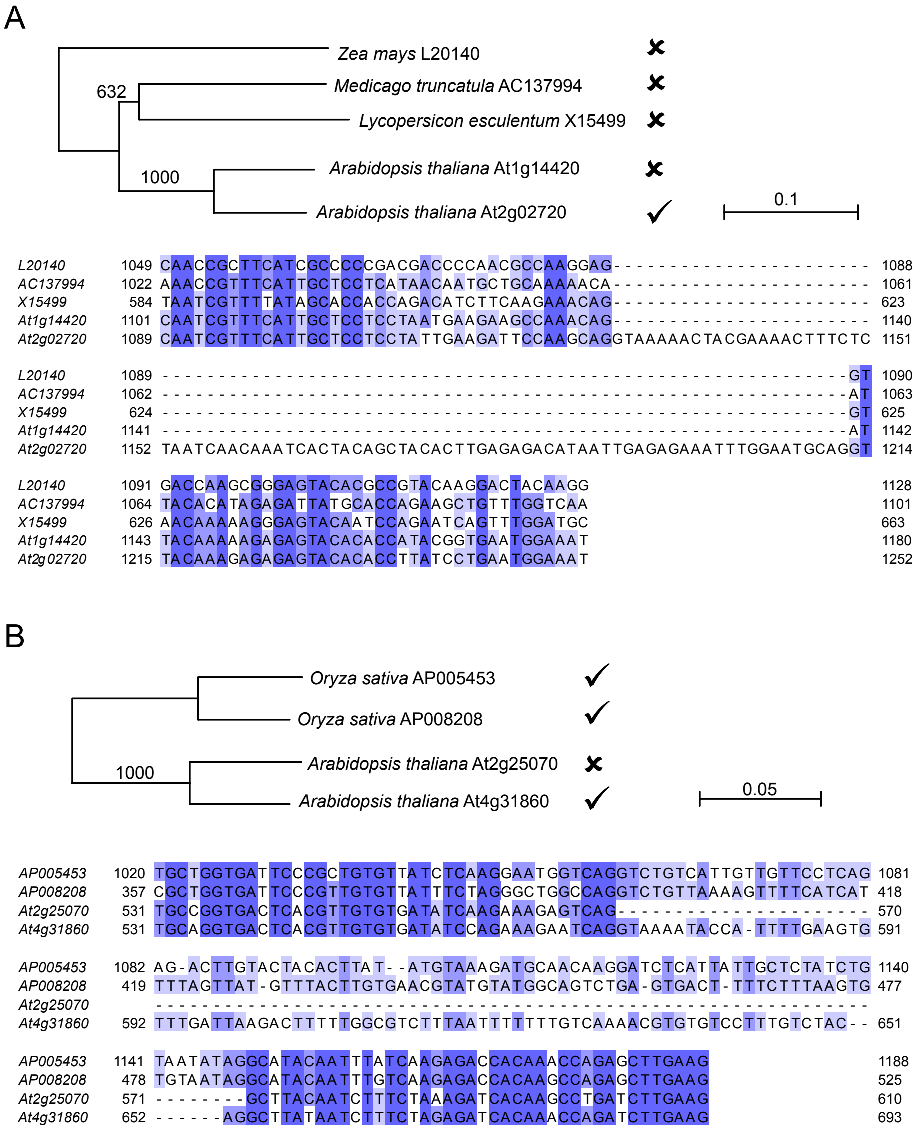

A
Zea mays L20140
Medicago truncatula AC137994
Lycopersicon esculentum X15499
Arabidopsis thaliana At1g14420
Arabidopsis thaliana At2g02720
632
1000
0.1
L20140 1049 CAACCGCTTCATCGCCCCCGACGACCCCAACGCCAAGGAG----------------------- 1088
AC137994 1022 AAACCGTTTCATTGCTCCTCATAACAATGCTGCAAAAACA----------------------- 1061
X15499 584 TAATCGTTTTATAGCACCACCAGACATCTTCAAGAAACAG----------------------- 623
At1g14420 1101 CAATCGTTTCATTGCTCCTCCTAATGAAGAAGCCAAACAG----------------------- 1140
At2g02720 1089 CAATCGTTTCATTGCTCCTCCTATTGAAGATTCCAAGCAGGTAAAAACTACGAAAACTTTCTC 1151
L20140 1089 ---GT 1090
AC137994 1062 ---AT 1063
X15499 624 ---GT 625
At1g14420 1141 ---AT 1142
At2g02720 1152 TAATCAACAAATCACTACAGCTACACTTGAGAGACATAATTGAGAGAAATTTGGAATGCAGGT 1214
L20140 1091 GACCAAGCGGGAGTACACGCCGTACAAGGACTACAAGG 1128
AC137994 1064 TACACATAGAGATTATGCACCAGAAGCTGTTTGGTCAA 1101
X15499 626 AACAAAAAGGGAGTACAATCCAGAATCAGTTTGGATGC 663
At1g14420 1143 TACAAAAAGAGAGTACACACCATACGGTGAATGGAAAT 1180
At2g02720 1215 TACAAAGAGAGAGTACACACCTTATCCTGAATGGAAAT 1252
B
Oryza sativa AP005453
Oryza sativa AP008208
Arabidopsis thaliana At2g25070
Arabidopsis thaliana At4g31860
1000
0.05
AP005453 1020 TGCTGGTGATTCCCGCTGTGTTATCTCAAGGAATGGTCAGGTCTGTCATTGTTGTTCCTCAG 1081
AP008208 357 CGCTGGTGATTCCCGTTGTGTTATTTCTAGGGCTGGCCAGGTCTGTTAAAAGTTTTCATCAT 418
At2g25070 531 TGCCGGTGACTCACGTTGTGTGATATCAAGAAAGAGTCAG---------------------- 570
At4g31860 531 TGCAGGTGACTCACGTTGTGTGATATCCAGAAAGAATCAGGTAAAATACCA-TTTTGAAGTG 591
AP005453 1082 AG-ACTTGTACTACACTTAT--ATGTAAAGATGCAACAAGGATCTCATTATTGCTCTATCTG 1140
AP008208 419 TTTAGTTAT-GTTTACTTGTGAACGTATGTATGGCAGTCTGA-GTGACT-TTTCTTTAAGTG 477
At2g25070 --
At4g31860 592 TTTGATTAAGACTTTTTGGCGTCTTTAATTTTTTTGTCAAAACGTGTGTCCTTTGTCTAC-- 651
AP005453 1141 TAATATAGGCATACAATTTATCAAGAGACCACAAACCAGAGCTTGAAG 1188
AP008208 478 TGTAATAGGCATACAATTTGTCAAGAGACCACAAGCCAGAGCTTGAAG 525
At2g25070 571 --------GCTTACAATCTTTCTAAAGATCACAAGCCTGATCTTGAAG 610
At4g31860 652 ------AGGCTTATAATCTTTCTAGAGATCACAAACCAGATCTTGAAG 693

Figure 6.1: A: Neighbor-joining tree and section of a multiple sequence alignment of the CDS, with the sequence of the examined intron added, from *A. thaliana* paralogues of the pectate lyase gene At2g02720. B: Neighbor-joining tree and section of a multiple sequence alignment of the CDS, with the examined intron sequences added, from *A. thaliana* paralogues of the protein phosphatase type 2C gene At2g25070. The intron is absent in this gene but present in its paralogue as well as in all identified plant homologues. The presence or absence of the intron in the species tree is indicated by ticks or crosses respectively. Bootstrap values (1000 replicates) are shown along the branches. The alignment shows the presence of an intron in this gene that is absent from all other identified plant homologues.

6.3.3 Intragenic Location of Intron Indels

Conflicting studies say that intron loss is (Mourier and Jeffares, 2003) or is not (Nielsen *et al.*, 2004) more prevalent in the 3' ends of genes in intron-poor genomes. A recent study by Lin and Zhang (2005) re-examined this question in many eukaryotic genomes, including *A. thaliana*, using a gene-by-gene method and found that all genomes analyzed display a significant 5' bias in the location of introns in genes irrespective of intron density (although *A. thaliana* showed the lowest bias).

When we examined the intron indels identified in this study using an intergene method similar to that of Nielsen *et al.* (2004), we did not find evidence for a bias in the intragenic location of gain and loss events, though they do appear to be more common in the middle of genes (see figure 6.2).

We also tested for bias in the location of intron indels using the intragene method of Lin and Zhang (2005) and found an excess of non-conserved introns in the 3' end of genes – only 189 genes display a 5' bias in the location of non-conserved introns compared with 273 genes that display a 3' bias; $P \leq 0.001$. The distribution of gained introns alone showed no significant bias. There were significantly more genes with a 3' bias of lost introns compared with a 5' bias (23 and 11 genes, respectively; $P \leq 0.05$). However, if we exclude genes with at least one ambiguous intron (*i.e.*, in a poorly aligned region or close to another intron), then there is no bias in the intragenic location of intron indels. The difference in the results from the two methods may be due to a greater robustness of the Lin and Zhang (2005) method to large variation in gene size because it only splits each gene into two location categories.

6.3.4 Relationship to Gene Expression and G+C Content

Previous studies have indicated that intron evolution is correlated with other genic and genomic features. Selection for transcriptional efficiency has led to the reduction in length (but not frequency) of introns in *A. thaliana* and other

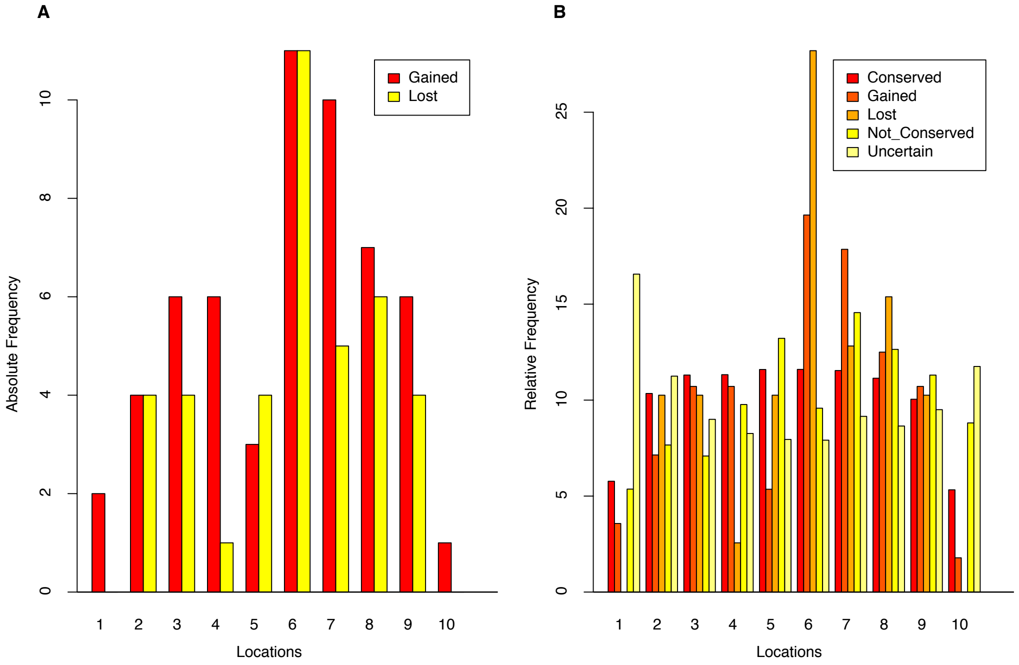


Figure 6.2: Relationship of intron fate to relative location within the gene. **A** Absolute frequencies of intron gain and loss events at different relative intra-genic locations. **B** Relative frequencies of conserved introns, gained introns, lost introns, other non-conserved introns, and unassigned introns (those with ambiguous alignment) within the gene.

eukaryotes (Castillo-Davis *et al.* 2002, Seoighe *et al.* 2005), and G +C rich regions of vertebrate genomes have shorter introns on average (Duret *et al.*, 1995).

We examined whether these phenomena known to influence intron length also influence intron gain and loss. We searched for evidence of a relationship between intron gain or loss and gene expression level based on microarray data and found no difference between genes with gained, lost, or conserved introns (figure 6.3).

Similarly, we found no significant difference in the G+C3 content of genes containing introns with different fates (figure 6.4).

6.3.5 Function of Genes experiencing Intron indels

We compared the function of pairs of genes with non-conserved introns with those that only contained conserved introns (and no ambiguous introns) using the GOslim Gene Ontology classifications from TAIR Berardini *et al.* (2004). A summary of the results is presented in table 6.2.

We did not consider gain and loss events separately because of low statistical power. Results uncorrected for multiple testing indicate that gene pairs that experienced at least one intron indel are enriched for gene ontology (GO) terms involving cytosol and hydrolase activity while transcription factor activity as well as unknown molecular function and biological processes are underrepresented; gene pairs that experienced at least 2 intron indels are enriched for other membranes, transport, and transporter activity; gene pairs that experienced at least 3 intron indels are enriched for nucleotide binding functions and signal transduction (all significant at the 1% level). When we repeated the simulations correcting for number of introns, the results were not significantly different. Coghlan and Wolfe (2004) previously found similar results in *Caenorhabditis* where many genes experiencing intron gains function in pre-mRNA processing.

There are 47 GOslim categories in this analysis. Because of multiple testing, if we consider each of these categories to be independent, we would expect just less than 2.5 categories to falsely appear significant at the 5% level and less than

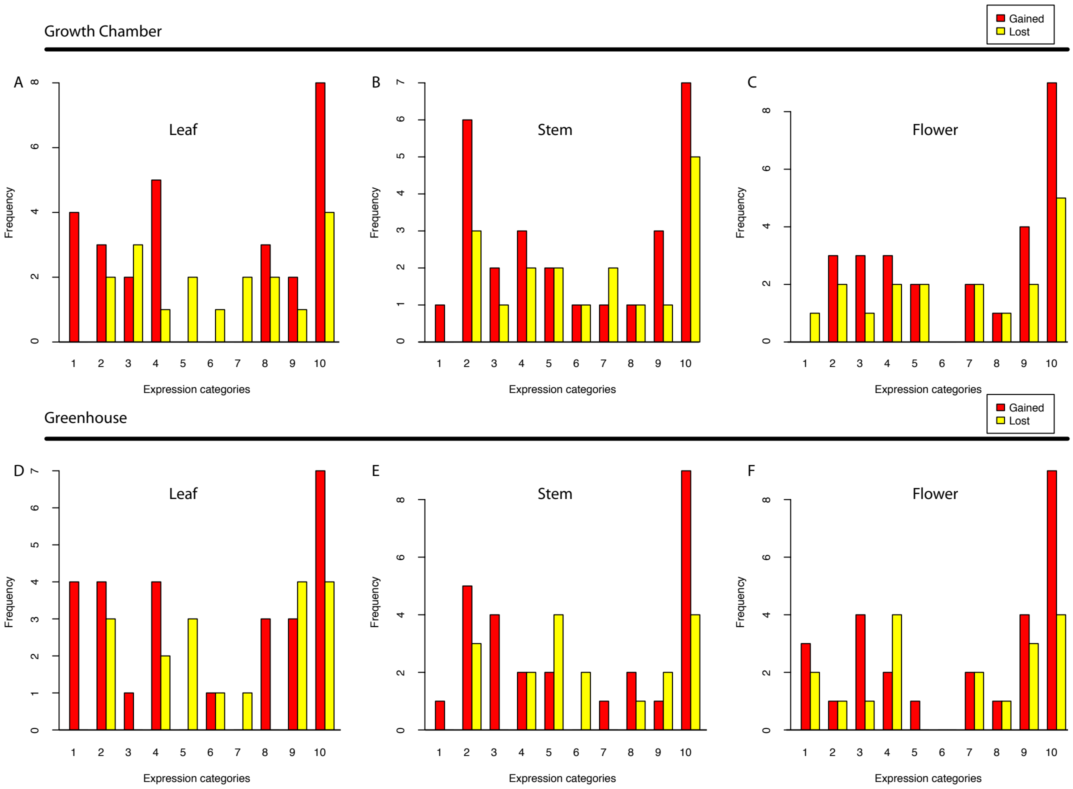


Figure 6.3: Relationship of intron gains and losses to the expression level of the gene. Expression levels of *A. thaliana* genes were binned into equal-sized categories. Panels A-C show growth chamber expression levels in leaf stem and flower respectively. D-F show greenhouse expression levels in leaf stem and flower respectively.

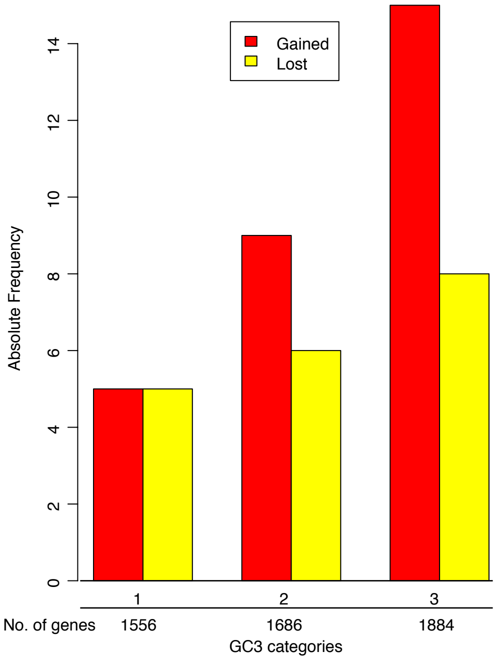


Figure 6.4: Relationship of intron gain and loss to GC3 content of the gene. All *A. thaliana* genes were ordered by GC3 content and binned into three equal sized GC3 categories. The frequencies of genes with gained or lost introns are shown. The total number of genes in our duplicated gene dataset that fall into each GC3 category are listed below the histogram.

Table 6.2: GO categories significantly over or under represented among gene pairs experiencing intron indels. (*) Indicates significance at the 5% level. (**) Indicates significance at the 1% level. Red shading indicates the result remained significant after Bonferroni correction for multiple testing at a 5% false positive rate and Benjamini-Hochberg (BH) correction to 5% false discovery rate (FDR). Yellow shading indicates the result remained significant after BH correction to 5% FDR.

GOslim Term	**Non-conserved introns ≥ 1** ($n = 281$)			**Non-conserved introns ≥ 2** ($n = 115$)			**Non-conserved introns ≥ 3** ($n = 37$)		
	Simulations		**Obs**	**Simulations**		**Obs**	**Simulations**		**Obs**
	Mean	**SD**		**Mean**	**SD**		**Mean**	**SD**	
Biological process unknown	122.29	8.20	🟥 96**	35.50	4.81	🟨 22**	11.43	2.80	7
Chloroplast	69.26	6.73	67	20.13	3.95	24	6.47	2.29	12*
Cytosol	6.17	2.19	🟨 13**	1.80	1.29	1	0.58	0.75	1
DNA and RNA binding	39.67	5.32	🟨 28*	11.52	3.12	6	3.71	1.81	2
DNA and RNA metabolism	5.34	2.05	4	1.55	1.20	2	0.50	0.70	2*
Hydrolase activity	46.87	5.74	🟨 63**	13.59	3.36	20	4.37	1.95	7
Kinase activity	30.23	4.73	🟨 42*	8.76	2.76	14	2.82	1.60	5
Molecular function unknown	104.82	7.83	🟥 82**	30.43	4.59	24	9.80	2.66	8
Nucleotide binding	31.26	4.79	🟨 41*	9.08	2.81	16*	2.92	1.62	8**
Nucleus	53.40	6.07	🟨 41*	15.51	3.56	15	4.99	2.06	7
Other membranes	109.13	7.92	116	31.67	4.65	🟨 44**	10.19	2.70	10
Response to stress	22.19	4.07	24	6.45	2.39	10	2.08	1.39	5*
Signal transduction	20.56	3.95	24	5.96	2.31	8	1.92	1.34	6**
Transcription	44.15	5.57	🟨 32*	12.84	3.29	9	4.13	1.90	4
Transcription factor activity	42.92	5.50	🟨 28**	12.49	3.24	7	4.02	1.88	3

Continued on next page

Table 6.2 – continued from previous page

GOslim Term	**Non-conserved introns $\geq$ 1** ($n = 281$)			**Non-conserved introns $\geq$ 2** ($n = 115$)			**Non-conserved introns $\geq$ 3** ($n = 37$)		
	Simulations		**Obs**	**Simulations**		**Obs**	**Simulations**		**Obs**
	Mean	**SD**		**Mean**	**SD**		**Mean**	**SD**	
Transferase activity	47.28	5.77	62*	13.71	3.38	19	4.41	1.95	6
Transport	34.32	5.00	46*	9.97	2.93	19**	3.22	1.70	6
Transporter activity	35.97	5.11	45	10.45	2.99	19**	3.37	1.73	4

0.5 categories to falsely appear significant at the 1% level. We observe more categories with P values ≤ 0.05 and ≤ 0.01, respectively, which indicates that most of these results are true positives but does not indicate which ones. When we correct for multiple testing using the Bonferroni correction, no GO terms are over-represented (table 6.2). However, Bonferroni correction is extremely strict, especially in cases where there may be some dependence between categories (as is the case with GO terms). When we use Benjamini-Hochberg correction (Benjamini and Hochberg, 1995), which aims to minimize the false discovery rate (FDR; i.e., the fraction of significant results that are actually false positives), genes with at least one intron indel are enriched for the terms *cytosol*, *hydrolase activity*, *kinase activity*, *nucleotide binding*, *transferase activity*, and *transport* at the 5% FDR level; genes with at least 2 intron indels are enriched for the terms *other membranes*, *transport*, and *transporter activity* also at the 5% FDR level (table 6.2). No terms remain significant for genes with at least 3 intron indels, which may be caused by low statistical power due to the small numbers of genes.

6.3.6 Search for Origins of Introns

The mechanisms of intron gain remain enigmatic (Fedorov *et al.*, 2003). Possible modes of intron origin almost all involve the incorporation of copies of other genomic DNA into a gene as an intron, be it directly through DNA duplication or indirectly through RNA intermediates (Roy and Gilbert, 2006). We searched the entire *A. thaliana* genome for DNA sequences with significant similarity to any of the non-conserved introns we identified, using BLAST and SSearch. We excluded self-hits, short hits (less than 50% of the query length), and any hits that were due to a large duplication (e.g., complete gene duplication, segmental chromosome duplication). We could successfully identify the origin of only one recently gained intron. The mechanically inducible TOUCH3 gene (TCH3; At2g41100) has gained an intron with respect to its paralogue the Calmodulin 3 gene (CAM3; At3g56800) that only contains one intron. TCH3 also contains one conserved intron and one intron in an ambiguous portion of the alignment with respect to CAM3. The sequence similarity search revealed that the new TCH3 intron is significantly similar to the conserved intron within TCH3 and to an intron in an adjacent paralogous gene in the chromosome (At2g41090). Inspection of the TCH3 gene sequence indicated that this intron was duplicated by a partial, internal gene duplication that also duplicated exonic sequence (figure 6.5 overleaf).

The 5 homologous sequences we identified in *Solanum tuberosum* (potato, 3 homologs), *Arachis hypogaea* (peanut), and *Oryza sativa* (rice) all resembled the CAM3 gene in gene structure and length.

The dotplot of the TCH3 gene against itself (figure 6.5 overleaf) indicates that much of the original gene was in fact duplicated twice (Sistrunk *et al.*, 1994), generating the 2 new introns (figure 6.6), but one of these (intron 2) was excluded by our alignment quality criteria during the assignment of conserved and non-conserved introns.

Interestingly, TCH3 has an alternative splice form (supported by cDNA evidence; GenBank NC_003071) that uses a pair of cryptic splice sites (AGGT) fortuitously present in the original duplicated gene segment as the ends of a new intron (figure 6.6 on page 193). This mechanism of intron gain was originally proposed over 15 years ago (Rogers, 1989).

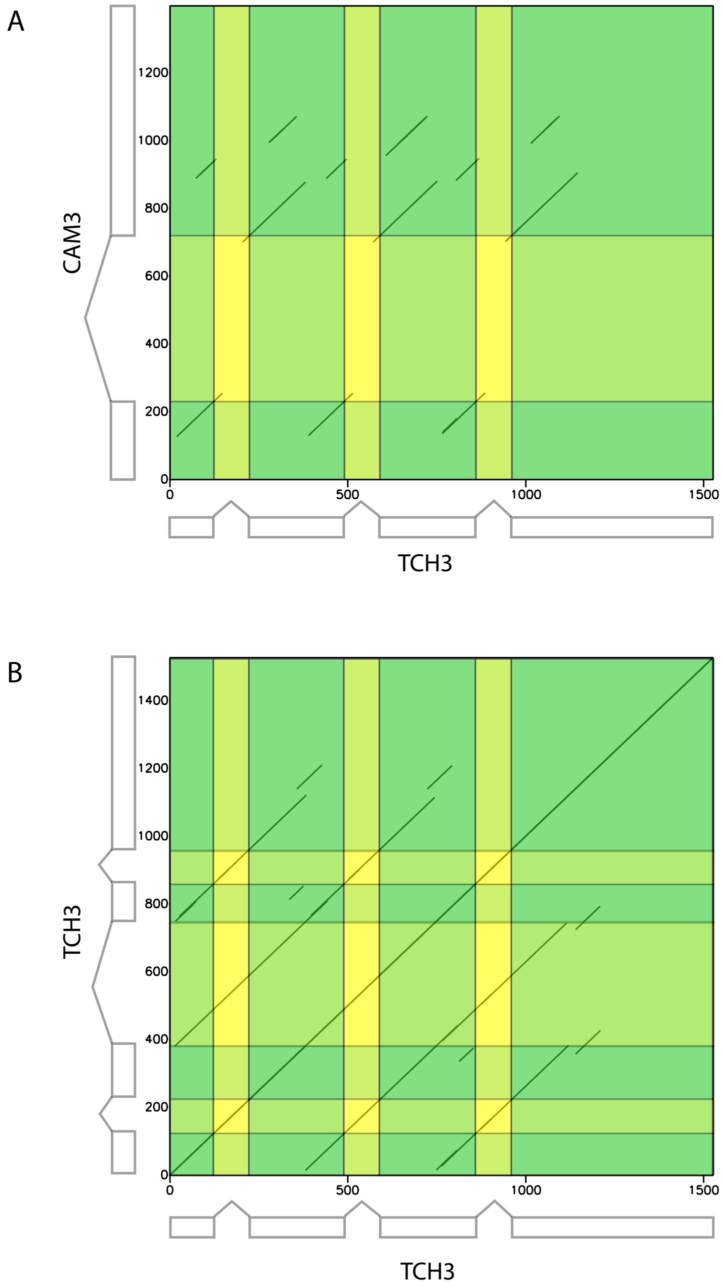


Figure 6.5: A: Dotplot representing the gene sequences of TCH3 (At2g41100) compared to the sequence of its paralogue CAM3 (At3g56800). B: TCH3 gene compared to itself. The alternatively spliced variant is indicated along the vertical axis. The plots were made with EMBOSS dotmatcher (see Methods 6.2). The exonic (including UTRs) and intronic regions are shown in green and yellow respectively and by a schematic representation of the spliced structure. The axes indicate the base position along the gene sequence for each gene.

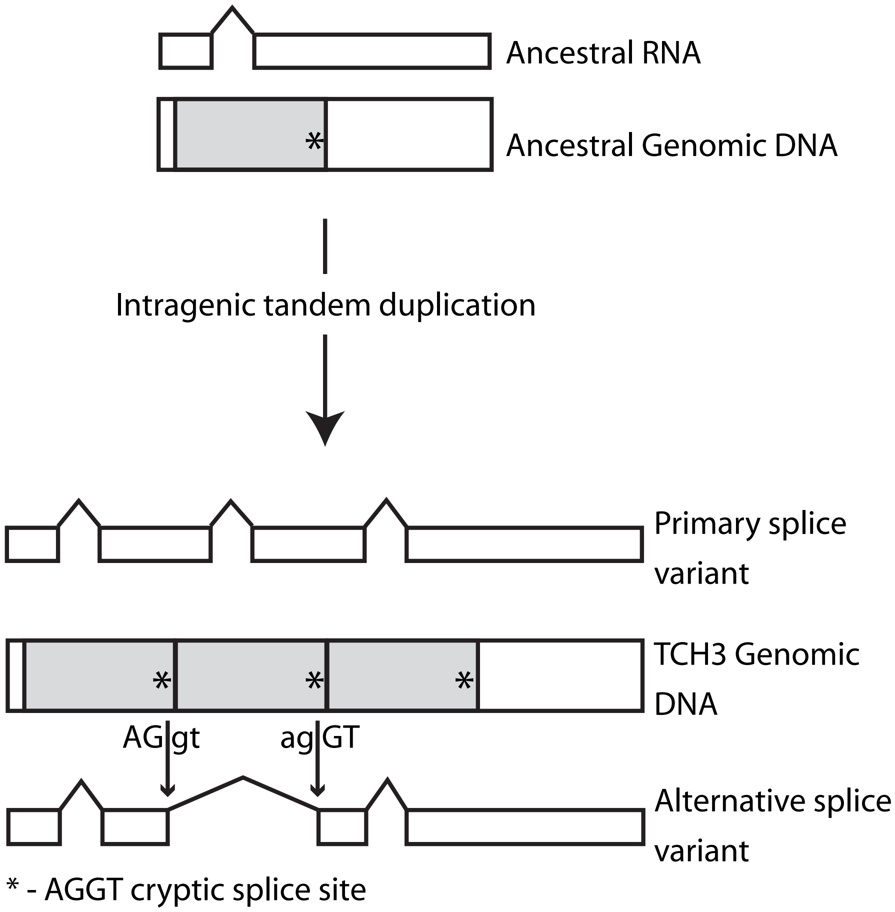


Figure 6.6: Schematic representation of the evolution of the TCH3 gene. The ancestral gene contained only one intron. Two intragenic duplications of an ancestral gene segment (shaded) copied an area of the ancestral gene twice generating two new introns in the primary splice variant of the modern TCH3 gene. The duplicated gene segment also included a cryptic splice site sequence AGGT – indicated by an asterisk (*)– close to its 3' end. A pair of cryptic splice sites is used as the boundaries of a new intron in the alternative splice variant of the modern TCH3 gene.

6.4 Discussion

The work reported here examines the dynamics of intron gain and loss on a much more recent scale than any previous studies. We observe a rate of gain and loss of introns of 2.7×10^{-3} to 9.1×10^{-4} events per intron site per million years (578 indel events out of 10,582 characterized intron locations in the 20 - 60 Myr since the genome duplication). This rate is higher than found in most previous studies stretching over broader evolutionary periods. If we extrapolate the amount of intron gain and loss to the whole data set (*i.e.*, 60% of non-conserved introns are gains), the rate of intron gain (2.0×10^{-12} to 5.9×10^{-12} gains per site per year; based on 2,873,004 possible insertion sites that pass alignment quality criteria) is similar to that found by Roy and Gilbert (2005a), and the rate of intron loss is orders of magnitude higher (4×10^{-6} to 1.2×10^{-5} events per year), although it is difficult to compare their study with ours because of methodological differences. The fact that this research focuses on relatively recently diverged genes gives greater power to detect intron gain and loss because over longer evolutionary periods, there is the opportunity for the gain and subsequent loss of an intron leading to underestimates of the number of events Roy and Gilbert (2006). For example, Roy *et al.* (2003) identified only 5 intron losses and no gains in 1500 human-mouse orthologues. Additionally, there may be some lineage-specific intron indel acceleration due to neutral drift to fixation of gain and loss polymorphisms facilitated by the tiny effective population size imposed by *A. thaliana*'s self-fertilization lifestyle.

Intron gain and loss in paralogous genes has been previously studied in a broad range of eukaryotic genomes (Babenko and Krylov 2004, Castillo-Davis *et al.* 2004, Qiu *et al.* 2004). Two studies examining this phenomenon in very old duplicate genes both found an excess of intron gain events (Babenko and Krylov 2004, Qiu *et al.* 2004). Analysis of introns in duplicated genes in *Plasmodium* malaria parasites of human and mouse indicated that intron indels are very frequent in paralogous genes, although they did not distinguish between gain and loss (Castillo-Davis *et al.*, 2004). One of the problems with these analyses lies in the estimation of the intron/exon structure at the time of duplication. The paralogous genes studied are likely to have been duplicated at widely different times – making rate estimation problematic – and by different mechanisms, including retrocopying of the gene via an mRNA intermediate that usually removes all introns from the gene. By contrast, the genes selected for analysis here were all duplicated at the same time and preserving gene structures.

All of these analyses of intron gain and loss find high rates of gene structure evolution in paralogues. Special features of paralogous genes that may give rise to higher intron flux include a possible contribution from subfunctionalization of alternative splice variants (Su *et al.*, 2006) which may involve changes in gene structure. However, this is more likely to involve the loss than the gain of an intron because it proposes the loss of alternative splicing by at least one of the duplicate genes. Some of the paralogous gene pairs in this study experienced multiple intron indels (table 6.1 and Supplementary table 2 at `http://mbe.oxfordjournals.org/cgi/content/full/msl017/DC1`), and it is not clear if there is something special about these genes. A GO term analysis indicates that the group of gene pairs with 2 or more non-conserved introns is enriched for functions involved in transport, transporter activity, and other membranes with respect to the entire group of paralogous genes. However, it is not clear why this should be the case.

Intron gain may be overestimated when there has been a parallel intron loss in the outgroup sequences. The Dollo Parsimony method we employed here does not attempt to correct for this, unlike likelihood methods. However, likelihood methods require an estimate of the rate of intron loss in order to estimate parallel loss events, and these estimates are not readily available for all lineages. A compromise has often been to assume constant rates on all lineages, which may not be biologically realistic.

Our analysis uncovered approximately 1.5 times more intron gains than losses during recent *A. thaliana* evolution (although the difference is not significant based on a χ^2 test). If the actual frequencies are equal in this data set, then that would imply a 9% chance of parallel loss in all outgroup sequences. If there is just a single outgroup from rice, this equates to a rate of intron loss of 4.5×10^{-4} per intron per million years (assuming a monocot-dicot divergence date of 200 MYA), which is comparable to the rate estimated by Roy and Gilbert (2005a). Where there are more numerous or more closely related outgroups, the rate of loss must be much higher to create this pattern of parallel intron loss. A tendency for parallel intron loss of particular introns (over random intron loss) has been observed in diverse *Caenorhabditis* genes and in the *White* gene of animals (Krzywinski and Besansky 2002, Cho *et al.* 2004). If this phenomenon holds in plant genomes or more generally if intron loss is more frequent in rice, then the parsimony method used here to infer intron gain and loss will be even more susceptible to the over-assignment of intron gains due to parallel loss in the outgroup.

A. thaliana has a famously small genome. One might have therefore naively predicted an excess of recent intron loss events, which we do not observe (though the amount of intron loss is high). However, the broad correlation between genome size and intron size in vertebrates (McLysaght *et al.*, 2000) is not generally apparent in plant genomes (Wendel *et al.*, 2002). This uncoupling of genome size and intron size is mirrored here by an uncoupling of genome reduction and intron loss. The yeast *Cryptococcus neoformans* has a similar uncoupling of these phenomena in its small, yet intron-dense, genome (Loftus *et al.*, 2005). In future research, it will be interesting to investigate whether the high rate of intron flux in paralogous genes can be related to subfunctionalization or neofunctionalization following gene duplication.

Chapter 7

Discussion

The evolution of the genome is a fascinating subject. Since the completion of the human genome we have learned a lot. With the addition of new genomes to the public databases, comparative genomics has become a powerful tool for the identification of conserved elements between genomes, as well as those that differ. The identification of conserved synteny between organisms allows for a more detailed study of the regions involved, and the increase in the quality of the available genomes allows for a better identification of those genes that may be unique to a certain lineage.

New genomes are completed at ever increasing speeds and already-available genomes are frequently updated with newly generated information and better annotation systems. This may make it difficult to generate results that are up to date with the most current data available, as in some cases by the time the results of a certain analysis are finished the initial data on which they are based is already old. Databases such as EnsEMBL and NCBI strive to maintain the data they contain available in specific forms maintaining these formats whenever possible when data are updated. This makes it increasingly useful to design analysis pipelines which can be updated with as little human intervention as possible in order to be able to keep up with the new genomic scale information as it becomes available.

During this study we have implemented an automated pipeline capable of downloading the required information and generating a set of synteny blocks for the comparison of any two species from the EnsEMBL database, as well as generating an initial classification of all protein coding genes contained in the two genomes. This pipeline will allow for a rapid survey of pairs of species in search of large areas of conserved synteny, and the result of these analyses can be also used as the basis for more comprehensive studies that include more species.

Our comparison of the genomes of human chimpanzee and macaque revealed a high degree of conservation as has been previously reported. This is reflected in the large size of many of the syntenic blocks we built between these genomes, as well as the large number of 1:1:1 orthologues shared between all three species. Most of the breaks in these synteny blocks were caused by the fragmentation of the chimpanzee and macaque genomes and so are not biologically relevant.

The gene classification we devised using a synteny block framework allowed us also to identify many genes that were present in only one of the two lineages. A strict examination of these candidates demonstrated that in most of the cases the absence of a clear orthologue at the expected location could be explained by the absence of sequence or annotation information in that area.

We used the human and chimpanzee comparison to search for possible cases of *de novo* gene formation from the exaptation of non-coding sequence within those genes that appeared to have no clear homology in the other hominid genome. In this search we identified nine cases for which we could not discard this possibility. There was detailed information available only for one of these genes, CLLU1 (Buhl *et al.*, 2006). From the available information we could not be certain to what extent the expression of this gene in normal circumstances may have any effect on the organism, as the background expression outside of the CLL cells, where it was identified, seems to be negligible. If all these candidates are indeed new genes this indicates the rate of non-coding sequence exaptation that produces new genes is quite high, with more than one case per Myr. However, due to the recent timing of these events their evolutionary fate is still uncertain.

The low background expression of the CLLU1 gene could be a feature of new genes formed by exaptation, which would make them even more difficult to detect from the expression data in organisms for which little EST data is available. This could be the case if the newly recruited promoter region has a low efficiency. A low expression level may benefit newly formed genes by allowing for the "testing of the gene" with expression levels that are less likely to damage the cell if the new gene interferes with existing processes in a deleterious manner. Indeed this seems to be the case for alternative splice variants that have originated from exonization of transposable elements (Sela *et al.*, 2007), which are usually present as a small fraction of the total transcripts from the gene.

If these genes prove to be beneficial they may be maintained by selection in the genome long enough to acquire mutations that allow the increase of the expression levels and become established in the lineage.

In many cases these exapted ORFs may come from highly degenerated retrocopies of genes. These inactivated copies may provide the possibility for drastic changes in the sequence of the originally encoded protein without any danger of the intermediate stages interacting in a detrimental manner with the original protein or its targets (Brosius, 1999). This could explain the similarity predicted at the structural level between the putative protein encoded by CLLU1 and the IL4 product (Buhl *et al.*, 2006).

We also searched for cases of alternatively spliced genes that have been duplicated in one of the two lineages and have subsequently lost different splice forms by a process of subfunctionalization. These events may be difficult to identify using many of the traditional approaches that assign gene orthology using only the longest transcript of each protein. This approach is probably valid for those organisms that have little alternative splicing, however in mammals there is increasing evidence that a large fraction of the genome undergoes alternative splicing. Although in most cases alternative transcripts are quite similar, this is not always the case as even when using some of the same exons the reading frame can be altered generating completely different protein sequences, as in the case of SPAG11 (Frohlich *et al.* 2001, Hall *et al.* 2007).

In our search, only one gene, SPAG11 was found that was, according to the available data, duplicated in human since the divergence from chimpanzee. However, although the two human copies show different sets of alternatively spliced forms according to the EnsEMBL annotation we were not able to find any evidence at the sequence level why any of these forms should not be capable of producing all the different alternative variants.

In all of these cases, the initial classification of the genes, and the search in human and chimpanzee for exaptations and subfunctionalized alternatively spliced genes, the main problem we encountered was the quality of the genomic data.

At the sequence level there are many gaps and regions of ambiguous sequence. At the annotation level we found several cases where genes have apparently been missed and others where they have been annotated incorrectly. Both of these problems may be responsible for many of the apparent gene content differences.

From the observation of these data we can conclude that the quality of the currently available data is insufficient to provide unambiguous data regarding the loss or absence of genes in these two primate species. This explains in a much simpler way the reason for the great number of gene gains and small number of losses reported for the human species in most of the existing genome comparison studies. Any other explanation would require a major change in the genome evolution mechanisms since the divergence of our species from the chimpanzee lineage 6 Mya.

Finally we examined the role of gain and loss at a lower level, examining the intron gain and loss events that have occurred in *Arabidopsis thaliana* since a major fraction of its genome was duplicated 20-60 Mya. We found that intron gain and loss events in this group of genes seem to occur with a much higher frequency than has been reported in other studies. In the subset of introns that could be assigned as gains or losses we found that a greater number of introns had been gained than lost. However most of the cases remained ambiguous in many cases due to the lack of sufficient genomic information from closely related species.

An effort is currently being made in order to obtain a draft sequence of many mammals distributed along the phylogenetic tree. This will give us a large amount of valuable information at a very good cost. However in order to fully identify the unique combination of features that separates one species from another, much greater quality genomes will be required, in both coverage and annotation.

Bibliography

AGI (2000). Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, **408**(6814), 796–815.

Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A., and Sorek, R. (2006). Transcription-mediated gene fusion in the human genome. *Genome Res*, **16**(1), 30–6.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2008). *Molecular Biology of the Cell*. Garland Science, 4th edition.

Alexeyenko, A., Lindberg, J., Prez-Bercoff, ., and Sonnhammer, E. L. (2006). Overview and comparison of ortholog databases. *Drug Discovery Today: Technologies*, **3**(22), 137–143.

Altschmied, J., Delfgaauw, J., Wilde, B., Duschl, J., Bouneau, L., Volff, J. N., and Schartl, M. (2002). Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *Genetics*, **161**(1), 259–67.

Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M., and Postlethwait, J. H. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science*, **282**(5394), 1711–4.

Antequera, F. (2003). Structure, function and evolution of cpg island promoters. *Cell Mol Life Sci*, **60**(8), 1647–58.

Aravind, L., Watanabe, H., Lipman, D. J., and Koonin, E. V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A*, **97**(21), 11319–24.

Asthana, S., Noble, W. S., Kryukov, G., Grant, C. E., Sunyaev, S., and Stamatoyannopoulos, J. A. (2007). Widely distributed noncoding purifying selection in the human genome. *Proc Natl Acad Sci U S A*, **104**(30), 12410–5.

Babenko, V. N. and Krylov, D. M. (2004). Comparative analysis of complete genomes reveals gene loss, acquisition and acceleration of evolutionary rates in metazoa, suggests a prevalence of evolution via gene acquisition and indicates that the evolutionary rates in animals tend to be conserved. *Nucleic Acids Res*, **32**(17), 5029–35.

Babushok, D. V., Ostertag, E. M., and Kazazian, H. H., J. (2007). Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol Life Sci*, **64**(5), 542–54.

Balding DJ., Bishop M., C. C. (2003). *Handbook of Statistical Genetics*, volume 1. Wiley, 2nd edition.

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., and Lander, E. S. (2002). Arachne: a whole-genome shotgun assembler. *Genome Res*, **12**(1), 177–89.

Begun, D. J., Lindfors, H. A., Kern, A. D., and Jones, C. D. (2007). Evidence for de novo evolution of testis-expressed genes in the drosophila yakuba/drosophila erecta clade. *Genetics*, **176**(2), 1131–7.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Berardini, T. Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L. A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D., and Rhee, S. Y. (2004). Functional annotation of the arabidopsis genome using controlled vocabularies. *Plant Physiol*, **135**(2), 745–55.

Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). *Biochemistry.* W. H. Freeman and Company, 5 edition.

Bernstein, R. M., Schluter, S. F., Bernstein, H., and Marchalonis, J. J. (1996). Primordial emergence of the recombination activating gene 1 (rag1): sequence of the complete shark gene indicates homology to microbial integrases. *Proc Natl Acad Sci U S A*, **93**(18), 9454–9.

Blanc, G. and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, **16**(7), 1667–78.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the arabidopsis genome. *Plant Cell*, **12**(7), 1093–101.

Blanc, G., Hokamp, K., and Wolfe, K. H. (2003). A recent polyploidy superimposed on older large-scale duplications in the arabidopsis genome. *Genome Res*, **13**(2), 137–44.

Blaustein, M., Pelisch, F., and Srebrow, A. (2007). Signals, pathways and splicing regulation. *Int J Biochem Cell Biol*, **39**(11), 2031–48.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol*, **7**(5), R43.

Bovee, D., Zhou, Y., Haugen, E., Wu, Z., Hayden, H. S., Gillett, W., Tuzun, E., Cooper, G. M., Sampas, N., Phelps, K., Levy, R., Morrison, V. A., Sprague, J., Jewett, D., Buckley, D., Subramaniam, S., Chang, J., Smith, D. R., Olson, M. V., Eichler, E. E., and Kaul, R. (2008). Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat Genet*, **40**(1), 96–101.

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**(6930), 433–8.

Brosius, J. (1999). Rnas from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, **238**(1), 115–34.

Buhl, A. M., Jurlander, J., Jorgensen, F. S., Ottesen, A. M., Cowland, J. B., Gjerdrum, L. M., Hansen, B. V., and Leffers, H. (2006). Identification of a gene on chromosome 12q22 uniquely overexpressed in chronic lymphocytic leukemia. *Blood*, **107**(7), 2904–11.

Burki, F. and Kaessmann, H. (2004). Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat Genet*, **36**(10), 1061–3.

Cai, J., Zhao, R., Jiang, H., and Wang, W. (2008). De novo origination of a new protein-coding gene in saccharomyces cerevisiae. *Genetics*, **179**(1), 487–96.

Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V., and Kondrashov, F. A. (2002). Selection for short introns in highly expressed genes. *Nat Genet*, **31**(4), 415–8.

Castillo-Davis, C. I., Bedford, T. B., and Hartl, D. L. (2004). Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites. *Mol Biol Evol*, **21**(7), 1422–7.

Castresana, J. (2002). Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high gc content. *Nucleic Acids Res*, **30**(8), 1751–6.

Castresana, J., Guigo, R., and Alba, M. M. (2004). Clustering of genes coding for dna binding proteins in a region of atypical evolution of the human genome. *J Mol Evol*, **59**(1), 72–9.

Cavalier-Smith, T. (2005). Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann Bot (Lond)*, **95**(1), 147–75.

Cavalier-Smith, T. (2006). Cell evolution and earth history: stasis and revolution. *Philos Trans R Soc Lond B Biol Sci*, **361**(1470), 969–1006.

Chen, L., DeVries, A. L., and Cheng, C. H. (1997). Convergent evolution of antifreeze glycoproteins in antarctic notothenioid fish and arctic cod. *Proc Natl Acad Sci U S A*, **94**(8), 3817–22.

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R. K., Paabo, S., Rocchi, M., and Eichler, E. E. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, **437**(7055), 88–93.

Cho, S., Jin, S. W., Cohen, A., and Ellis, R. E. (2004). A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. *Genome Res*, **14**(7), 1207–20.

Coghlan, A. and Wolfe, K. H. (2004). Origins of recently gained introns in caenorhabditis. *Proc Natl Acad Sci U S A*, **101**(31), 11362–7.

Cresko, W. A., Yan, Y. L., Baltrus, D. A., Amores, A., Singer, A., Rodriguez-Mari, A., and Postlethwait, J. H. (2003). Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. *Dev Dyn*, **228**(3), 480–9.

CSAC (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**(7055), 69–87.

Cusack, B. P. and Wolfe, K. H. (2007). When gene marriages don't work out: divorce by subfunctionalization. *Trends Genet*, **23**(6), 270–2.

Delaye, L., Deluna, A., Lazcano, A., and Becerra, A. (2008). The origin of a novel gene through overprinting in escherichia coli. *BMC Evol Biol*, **8**(31), 31.

Delneri, D., Colson, I., Grammenoudi, S., Roberts, I. N., Louis, E. J., and Oliver, S. G. (2003). Engineering evolution to study speciation in yeasts. *Nature*, **422**(6927), 68–72.

Demuth, J. P., Bie, T. D., Stajich, J. E., Cristianini, N., and Hahn, M. W. (2006). The evolution of mammalian gene families. *PLoS ONE*, **1**(1), e85.

Deutsch, M. and Long, M. (1999). Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res*, **27**(15), 3219–28.

Devos, K. M., Brown, J. K., and Bennetzen, J. L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in arabidopsis. *Genome Res*, **12**(7), 1075–9.

Dickson, L., Huang, H. R., Liu, L., Matsuura, M., Lambowitz, A. M., and Perlman, P. S. (2001). Retrotransposition of a yeast group ii intron occurs by reverse splicing directly into ectopic dna sites. *Proc Natl Acad Sci U S A*, **98**(23), 13207–12.

Duret, L., Mouchiroud, D., and Gautier, C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in gc-rich isochores. *J Mol Evol*, **40**(3), 308–17.

Eckhart, L., Ballaun, C., Hermann, M., VandeBerg, J. L., Sipos, W., Uthman, A., Fischer, H., and Tschachler, E. (2008). Identification of novel mammalian caspases reveals an important role of gene loss in shaping the human caspase repertoire. *Mol Biol Evol*, **25**(5), 831–41.

Eichler, E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet*, **17**(11), 661–9.

Elango, N., Thomas, J. W., and Yi, S. V. (2006). Variable molecular clocks in hominoids. *Proc Natl Acad Sci U S A*, **103**(5), 1370–5.

Fedorov, A., Roy, S., Fedorova, L., and Gilbert, W. (2003). Mystery of intron gain. *Genome Res*, **13**(10), 2236–41.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**(4), 1531–45.

Frohlich, O., Po, C., and Young, L. G. (2001). Organization of the human gene encoding the epididymis-specific ep2 protein variants and its relationship to defensin genes. *Biol Reprod*, **64**(4), 1072–9.

Gallardo, M. H., Bickham, J. W., Kausel, G., Kohler, N., and Honeycutt, R. L. (2003). Gradual and quantum genome size shifts in the hystricognath rodents. *J Evol Biol*, **16**(1), 163–9.

Gerber, A., O'Connell, M. A., and Keller, W. (1997). Two forms of human double-stranded rna-specific editase 1 (hred1) generated by the insertion of an alu cassette. *Rna*, **3**(5), 453–63.

Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., Remington, K. A., Strausberg, R. L., Venter, J. C., Wilson, R. K., Batzer, M. A., Bustamante, C. D., Eichler, E. E., Hahn, M. W., Hardison, R. C., Makova, K. D., Miller, W., Milosavljevic, A., Palermo, R. E., Siepel, A., Sikela, J. M., Attaway, T., Bell, S., Bernard, K. E., Buhay, C. J., Chandrabose, M. N., Dao, M., Davis, C., Delehaunty, K. D., Ding, Y., Dinh, H. H., Dugan-Rocha, S., Fulton, L. A., Gabisi, R. A., Garner, T. T., Godfrey, J., Hawes, A. C., Hernandez, J., Hines, S., Holder, M., Hume, J., Jhangiani, S. N., Joshi, V., Khan, Z. M., Kirkness, E. F., Cree, A., Fowler, R. G., Lee, S., Lewis, L. R., Li, Z., Liu, Y. S., Moore, S. M., Muzny, D., Nazareth, L. V., Ngo, D. N., Okwuonu, G. O., Pai, G., Parker, D., Paul, H. A., Pfannkoch, C., Pohl, C. S., Rogers, Y. H., Ruiz, S. J., Sabo, A., Santibanez, J., Schneider, B. W., Smith, S. M., Sodergren, E., Svatek, A. F., Utterback, T. R., Vattathil, S., Warren, W., White, C. S., Chinwalla, A. T., Feng, Y., Halpern, A. L., Hillier, L. W., Huang, X., Minx, P., Nelson, J. O., Pepin, K. H., Qin, X., Sutton, G. G., Venter, E., Walenz, B. P., Wallis, J. W., Worley, K. C., Yang, S. P., Jones, S. M., Marra, M. A., Rocchi, M., Schein, J. E., Baertsch, R., Clarke, L., Csuros, M., Glasscock, J., Harris, R. A., Havlak, P., Jackson, A. R., Jiang, H., *et al.* (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**(5822), 222–34.

Gilad, Y., Man, O., and Glusman, G. (2005). A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res*, **15**(2), 224–30.

Gilson, P. R. and McFadden, G. I. (1996). The miniaturized nuclear genome of eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns. *Proc Natl Acad Sci U S A*, **93**(15), 7737–42.

Glazko, G. V. and Nei, M. (2003). Estimation of divergence times for major lineages of primate species. *Mol Biol Evol*, **20**(3), 424–34.

Goodstadt, L. and Ponting, C. P. (2006). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol*, **2**(9), e133.

Goodstadt, L., Heger, A., Webber, C., and Ponting, C. P. (2007). An analysis of the gene complement of a marsupial, monodelphis domestica: Evolution of lineage-specific genes and giant chromosomes. *Genome Res*, **10**, 10.

Graur, D. and Li, W.-H. (2000). *Fundamentals of Molecular Evolution (Second Edition)*. Sinauer.

Green, P. (1997). Against a whole-genome shotgun. *Genome Res*, **7**(5), 410–7.

Gregory, T. R. (2005). Synergy between sequence and size in large-scale genomics. *Nat Rev Genet*, **6**(9), 699–708.

Grimwood, J., Gordon, L. A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D., Couronne, O., Tran-Gyamfi, M., Aerts, A., Altherr, M., Ashworth, L., Bajorek, E., Black, S., Branscomb, E., Caenepeel, S., Carrano, A., Caoile, C., Chan, Y. M., Christensen, M., Cleland, C. A., Copeland, A., Dalin, E., Dehal, P., Denys, M., Detter, J. C., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Georgescu, A. M., Glavina, T., Gomez, M., Gonzales, E., Groza, M., Hammon, N., Hawkins, T., Haydu, L., Ho, I., Huang, W., Israni, S., Jett, J., Kadner, K., Kimball, H., Kobayashi, A., Larionov, V., Leem, S. H., Lopez, F., Lou, Y., Lowry, S., Malfatti, S., Martinez, D., McCready, P., Medina, C., Morgan, J., Nelson, K., Nolan, M., Ovcharenko, I., Pitluck, S., Pollard, M., Popkie, A. P., Predki, P., Quan, G., Ramirez, L., Rash, S., Retterer, J., Rodriguez, A., Rogers, S., Salamov, A., Salazar, A., She, X., Smith, D., Slezak, T., Solovyev, V., Thayer, N., Tice, H., Tsai, M., Ustaszewska, A., Vo, N., Wagner, M., Wheeler, J., Wu, K., Xie, G., Yang, J., Dubchak, I., Furey, T. S., DeJong, P., Dickson, M., Gordon, D., Eichler, E. E., Pennacchio, L. A., Richardson, P., Stubbs, L., Rokhsar, D. S., Myers, R. M., Rubin, E. M., and Lucas, S. M. (2004). The dna sequence and biology of human chromosome 19. *Nature*, **428**(6982), 529–35.

Guigo, R., Flicek, P., Abril, J. F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V. B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T. R., Harrow, J., Hubbard, T., Lewis, S. E., and Reese, M. G. (2006). Egasp: the human encode genome annotation assessment project. *Genome Biol*, **7 Suppl 1**(1), S2 1–31.

Gure, A. O., Wei, I. J., Old, L. J., and Chen, Y. T. (2002). The ssx gene family: characterization of 9 complete genes. *Int J Cancer*, **101**(5), 448–53.

Hahn, M. W., Demuth, J. P., and Han, S. G. (2007). Accelerated rate of gene gain and loss in primates. *Genetics*, **18**, 18.

Hall, S. H., Yenugu, S., Radhakrishnan, Y., Avellar, M. C., Petrusz, P., and French, F. S. (2007). Characterization and functions of beta defensins in the epididymis. *Asian J Androl*, **9**(4), 453–62.

Harder, J., Siebert, R., Zhang, Y., Matthiesen, P., Christophers, E., Schlegelberger, B., and Schroder, J. M. (1997). Mapping of the gene encoding human beta-defensin-2 (defb2) to chromosome region 8p22-p23.1. *Genomics*, **46**(3), 472–5.

Hayasaka, K., Gojobori, T., and Horai, S. (1988). Molecular phylogeny and evolution of primate mitochondrial dna. *Mol Biol Evol*, **5**(6), 626–44.

Hedges, S. B., Blair, J. E., Venturi, M. L., and Shoe, J. L. (2004). A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol*, **4**(2), 2.

Hiebl, I., Weber, R. E., Schneeganss, D., Kosters, J., and Braunitzer, G. (1988). High-altitude respiration of birds. structural adaptations in the major and minor hemoglobin components of adult ruppell's griffon (gyps rueppellii, aegypiinae): a new molecular pattern for hypoxic tolerance. *Biol Chem Hoppe Seyler*, **369**(4), 217–32.

Hittinger, C. T. and Carroll, S. B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, **449**(7163), 677–681.

Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat Med*, **9**(7), 811–8.

Hoegg, S., Boore, J. L., Kuehl, J. V., and Meyer, A. (2007). Comparative phylogenomic analyses of teleost fish hox gene clusters: lessons from the cichlid fish astatotilapia burtoni. *BMC Genomics*, **8**(317), 317.

Huang, X., Wang, J., Aluru, S., Yang, S. P., and Hillier, L. (2003). Pcap: a whole-genome assembly program. *Genome Res*, **13**(9), 2164–70.

Hubbard, T. J., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., and Birney, E. (2007). Ensembl 2007. *Nucleic Acids Res*, **35**(Database issue), D610–7.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldon, T. (2007). The human phylome. *Genome Biol*, **8**(6), R109.

Hughes, A. L. and Friedman, R. (2004). Differential loss of ancestral gene families as a source of genomic divergence in animals. *Proc Biol Sci*, **271 Suppl 3**(271), S107–9.

Hughes, A. L. and Hughes, M. K. (1995). Small genomes for better flyers. *Nature*, **377**(6548), 391.

Huynen, M. A. and Bork, P. (1998). Measuring genome evolution. *Proc Natl Acad Sci U S A*, **95**(11), 5849–56.

IHGSC (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011), 931–45.

Imai, S., Sasaki, T., Shimizu, A., Asakawa, S., Hori, H., and Shimizu, N. (2007). The genome size evolution of medaka (oryzias latipes) and fugu (takifugu rubripes). *Genes Genet Syst*, **82**(2), 135–44.

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biemont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J. P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J. N., Guigo, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quetier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J., and Roest Crollius, H. (2004). Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**(7011), 946–57.

Jaillon, O., Bouhouche, K., Gout, J. F., Aury, J. M., Noel, B., Saudemont, B., Nowacki, M., Serrano, V., Porcel, B. M., Segurens, B., Le Mouel, A., Lepere, G., Schachter, V., Betermier, M., Cohen, J., Wincker, P., Sperling, L., Duret, L., and Meyer, E. (2008). Translational control of intron splicing in eukaryotes. *Nature*, **451**(7176), 359–62.

Javaud, C., Dupuy, F., Maftah, A., Julien, R., and Petit, J. M. (2003). The fucosyltransferase gene family: an amazing summary of the underlying mechanisms of gene evolution. *Genetica*, **118**(2-3), 157–70.

Jeffares, D. C., Mourier, T., and Penny, D. (2006). The biology of intron gain and loss. *Trends Genet*, **22**(1), 16–22.

Johnson, K. R., Wright, J. E., J., and May, B. (1987). Linkage relationships reflecting ancestral tetraploidy in salmonid fish. *Genetics*, **116**(4), 579–91.

Kapitonov, V. V. and Jurka, J. (2005). Rag1 core and v(d)j recombination signal sequences were derived from transib transposons. *PLoS Biol*, **3**(6), e181.

Katju, V. and Lynch, M. (2006). On the formation of novel genes by duplication in the caenorhabditis elegans genome. *Mol Biol Evol*, **23**(5), 1056–1067.

Keese, P. K. and Gibbs, A. (1992). Origins of genes: "big bang" or continuous creation? *Proc Natl Acad Sci U S A*, **89**(20), 9489–93.

Kehrer-Sawatzki, H. and Cooper, D. N. (2007a). Structural divergence between the human and chimpanzee genomes. *Hum Genet*, **120**(6), 759–78.

Kehrer-Sawatzki, H. and Cooper, D. N. (2007b). Understanding the recent evolution of the human genome: insights from human-chimpanzee genome comparisons. *Hum Mutat*, **28**(2), 99–130.

Kent, W. J. and Haussler, D. (2001). Assembly of the working draft of the human genome with gigassembler. *Genome Res*, **11**(9), 1541–8.

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, **35**(1), 125–31.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J. M., and Beslon, G. (2007). A long-term evolutionary pressure on the amount of non-coding dna. *Mol Biol Evol*, **19**, 19.

Knowles, D. G. and McLysaght, A. (2006). High rate of recent intron gain and loss in simultaneously duplicated arabidopsis genes. *Mol Biol Evol*, **23**, 23.

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**(2), 187–208.

Krzywinski, J. and Besansky, N. J. (2002). Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. *Mol Biol Evol*, **19**(3), 362–6.

Kummerfeld, S. K. and Teichmann, S. A. (2005). Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet*, **21**(1), 25–30.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Lauer, P., Schneider, S. S., and Gnirke, A. (1998). Construction and validation of yeast artificial chromosome contig maps by reca-assisted restriction endonuclease cleavage. *Proc Natl Acad Sci U S A*, **95**(19), 11318–23.

Letunic, I., Copley, R. R., and Bork, P. (2002). Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet*, **11**(13), 1561–7.

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., and Begun, D. J. (2006). Novel genes derived from noncoding dna in drosophila melanogaster are frequently x-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A*, **103**(26), 9935–9.

Lewontin, R. (2000). *The triple helix*. Harvard University Press.

Li, W. H. (1993). So, what about the molecular clock hypothesis? *Curr Opin Genet Dev*, **3**(6), 896–901.

Li, X., Li, H. P., Amsler, K., Hyink, D., Wilson, P. D., and Burrow, C. R. (2002). Prkx, a phylogenetically and functionally distinct camp-dependent protein kinase, activates renal epithelial cell migration and morphogenesis. *Proc Natl Acad Sci U S A*, **99**(14), 9260–5.

Lin, H., Zhu, W., Silva, J. C., Gu, X., and Buell, C. R. (2006). Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol*, **7**(5), R41.

Lin, K. and Zhang, D. Y. (2005). The excess of 5' introns in eukaryotic genomes. *Nucleic Acids Res*, **33**(20), 6522–7.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., deJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C., and Lander, E. S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**(7069), 803–819.

Liu, X. Z., Li, S. L., Jing, H., Liang, Y. H., Hua, Z. Q., and Lu, G. Y. (2001). Avian haemoglobins and structural basis of high affinity for oxygen: structure of bar-headed goose aquomet haemoglobin. *Acta Crystallogr D Biol Crystallogr*, **57**(Pt 6), 775–83.

Loftus, B. J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I. J., Fraser, J. A., Allen, J. E., Bosdet, I. E., Brent, M. R., Chiu, R., Doering, T. L., Donlin, M. J., D'Souza, C. A., Fox, D. S., Grinberg, V., Fu, J., Fukushima, M., Haas, B. J., Huang, J. C., Janbon, G., Jones, S. J., Koo, H. L., Krzywinski, M. I., Kwon-Chung, J. K., Lengeler, K. B., Maiti, R., Marra, M. A., Marra, R. E., Mathewson, C. A., Mitchell, T. G., Pertea, M., Riggs, F. R., Salzberg, S. L., Schein, J. E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C. A., Suh, B. B., Tenney, A., Utterback, T. R., Wickes, B. L., Wortman, J. R., Wye, N. H., Kronstad, J. W., Lodge, J. K., Heitman, J., Davis, R. W., Fraser, C. M., and Hyman, R. W. (2005). The genome of the basidiomycetous yeast and human pathogen cryptococcus neoformans. *Science*, **307**(5713), 1321–4.

Long, M. and Langley, C. H. (1993). Natural selection and the origin of jingwei, a chimeric processed functional gene in drosophila. *Science*, **260**(5104), 91–5.

Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat Rev Genet*, **4**(11), 865–75.

Lopez-Bigas, N., De, S., and Teichmann, S. A. (2008). Functional protein divergence in the evolution of homo sapiens. *Genome Biol*, **9**(2), R33.

Lynch, M. (2005). Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol*, **23**, 23.

Lynch, M. (2006). The origins of eukaryotic gene structure. *Mol Biol Evol*, **23**(2), 450–68.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**(5494), 1151–5.

Lynch, M. and Richardson, A. O. (2002). The evolution of spliceosomal introns. *Curr Opin Genet Dev*, **12**(6), 701–10.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*, **102**(15), 5454–9.

Makalowska, I., Lin, C. F., and Hernandez, K. (2007). Birth and death of gene overlaps in vertebrates. *BMC Evol Biol*, **7**(193), 193.

Martin, W. and Koonin, E. V. (2006). Introns and the origin of nucleus-cytosol compartmentalization. *Nature*, **440**(7080), 41–5.

Maston, G. A. and Ruvolo, M. (2002). Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. *Mol Biol Evol*, **19**(3), 320–35.

McLysaght, A., Enright, A. J., Skrabanek, L., and Wolfe, K. H. (2000). Estimation of synteny conservation and genome compaction between pufferfish (fugu) and human. *Yeast*, **17**(1), 22–36.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**(5670), 581–4.

Mirsky, A. E. and Ris, H. (1951). The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J Gen Physiol*, **34**(4), 451–62.

Mockler, T. C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S. E., and Ecker, J. R. (2005). Applications of dna tiling arrays for whole-genome analysis. *Genomics*, **85**(1), 1–15.

Mourier, T. and Jeffares, D. C. (2003). Eukaryotic intron loss. *Science*, **300**(5624), 1393.

Muller, H. J. (1925). Why polyploidy is rarer in animals than in plants. *The American Naturalist*, **59**(663), 346–353.

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W., and Springer, M. S. (2001). Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science*, **294**(5550), 2348–51.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A., Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Skow, L. C., Welge, M., Womack, J. E., O'Brien S, J., Pevzner, P. A., and Lewin, H. A. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, **309**(5734), 613–7.

Nei, M. and Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*, **39**, 121–52.

Nei, M., Xu, P., and Glazko, G. (2001). Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc Natl Acad Sci U S A*, **98**(5), 2497–502.

Nguyen, H. D., Yoshihama, M., and Kenmochi, N. (2005). New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput Biol*, **1**(7), e79.

Nielsen, C. B., Friedman, B., Birren, B., Burge, C. B., and Galagan, J. E. (2004). Patterns of intron gain and loss in fungi. *PLoS Biol*, **2**(12), e422.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nat Rev Genet*, **8**(11), 857–68.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1), 205–17.

Oda, M., Satta, Y., Takenaka, O., and Takahata, N. (2002). Loss of urate oxidase activity in hominoids and its evolutionary implications. *Mol Biol Evol*, **19**(5), 640–53.

Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag.

Oliver, M. J., Petrov, D., Ackerly, D., Falkowski, P., and Schofield, O. M. (2007). The mode and tempo of genome size evolution in eukaryotes. *Genome Res*, **9**, 9.

Olson, M. V. (1999). When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet*, **64**(1), 18–23.

Oort, P. J., Warden, C. H., Baumann, T. K., Knotts, T. A., and Adams, S. H. (2007). Characterization of tusc5, an adipocyte gene co-expressed in peripheral neurons. *Mol Cell Endocrinol*, **276**(1-2), 24–35.

Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, E. T., Castelo, R., Thomson, T. M., Antonarakis, S. E., and Guigo, R. (2006). Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res*, **16**(1), 37–44.

Paterson, A. H., Bowers, J. E., Burow, M. D., Draye, X., Elsik, C. G., Jiang, C. X., Katsar, C. S., Lan, T. H., Lin, Y. R., Ming, R., and Wright, R. J. (2000). Comparative genomics of plant chromosomes. *Plant Cell*, **12**(9), 1523–40.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol*, **266**, 227–58.

Pevzner, P. and Tesler, G. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A*, **100**(13), 7672–7.

Pheasant, M. and Mattick, J. S. (2007). Raising the estimate of functional human sequences. *Genome Res*, **9**, 9.

Potrzebowski, L., Vinckenbosch, N., Marques, A. C., Chalmel, F., Jegou, B., and Kaessmann, H. (2008). Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol*, **6**(4), e80.

Prince, V. E. and Pickett, F. B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet*, **3**(11), 827–37.

Purvis, A. (1995). A composite estimate of primate phylogeny. *Philos Trans R Soc Lond B Biol Sci*, **348**(1326), 405–21.

Qiu, W. G., Schisler, N., and Stoltzfus, A. (2004). The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol*, **21**(7), 1252–63.

Raaum, R. L., Sterner, K. N., Noviello, C. M., Stewart, C. B., and Disotell, T. R. (2005). Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear dna evidence. *J Hum Evol*, **48**(3), 237–57.

Reyes, A., Gissi, C., Catzeflis, F., Nevo, E., Pesole, G., and Saccone, C. (2004). Congruent mammalian trees from mitochondrial and nuclear genes using bayesian methods. *Mol Biol Evol*, **21**(2), 397–403.

Reymond, A., Camargo, A. A., Deutsch, S., Stevenson, B. J., Parmigiani, R. B., Ucla, C., Bettoni, F., Rossier, C., Lyle, R., Guipponi, M., de Souza, S., Iseli, C., Jongeneel, C. V., Bucher, P., Simpson, A. J., and Antonarakis, S. E. (2002). Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics*, **79**(6), 824–32.

Rodriguez-Trelles, F., Tarrio, R., and Ayala, F. J. (2006). Origins and evolution of spliceosomal introns. *Annu Rev Genet*, **40**, 47–76.

Rogers, J. H. (1989). How were introns inserted into nuclear genes? *Trends Genet*, **5**(7), 213–6.

Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G., and Koonin, E. V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*, **13**(17), 1512–7.

Rogozin, I. B., Sverdlov, A. V., Babenko, V. N., and Koonin, E. V. (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform*, **6**(2), 118–34.

Roy, S. W. and Gilbert, W. (2005a). Complex early genes. *Proc Natl Acad Sci U S A*, **102**(6), 1986–91.

Roy, S. W. and Gilbert, W. (2005b). Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc Natl Acad Sci U S A*, **102**(16), 5773–8.

Roy, S. W. and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet*, **7**(3), 211–21.

Roy, S. W., Fedorov, A., and Gilbert, W. (2003). Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci U S A*, **100**(12), 7158–62.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S. (2006). Positive natural selection in the human lineage. *Science*, **312**(5780), 1614–20.

Samonte, R. V. and Eichler, E. E. (2002). Segmental duplications and the evolution of the primate genome. *Nat Rev Genet*, **3**(1), 65–72.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, **74**(12), 5463–7.

Scannell, D. R., Frank, A. C., Conant, G. C., Byrne, K. P., Woolfit, M., and Wolfe, K. H. (2007). Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci U S A*, **104**(20), 8397–402.

Schiebel, K., Winkelmann, M., Mertz, A., Xu, X., Page, D. C., Weil, D., Petit, C., and Rappold, G. A. (1997). Abnormal xy interchange between a novel isolated protein kinase gene, prky, and its homologue, prkx, accounts for one third of all (y+)xx males and (y-)xy females. *Hum Mol Genet*, **6**(11), 1985–9.

Schmidt, E. E. and Schibler, U. (1995). High accumulation of components of the rna polymerase ii transcription machinery in rodent spermatids. *Development*, **121**(8), 2373–83.

Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y. M., Denys, M., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Gomez, M., Gonzales, E., Haydu, L., Lopez, F., Ramirez, L., Retterer, J., Rodriguez, A., Rogers, S., Salazar, A., Tsai, M., and Myers, R. M. (2004). Quality assessment of the human genome sequence. *Nature*, **429**(6990), 365–8.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). Smart, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A*, **95**(11), 5857–64.

Schwager, E. E., Schoppmeier, M., Pechmann, M., and Damen, W. G. (2007). Duplicated hox genes in the spider cupiennius salei. *Front Zool*, **4**(1), 10.

Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). Pipmaker–a web server for aligning two genomic dna sequences. *Genome Res*, **10**(4), 577–86.

Sela, N., Mersch, B., Gal-Mark, N., Lev-Maor, G., Hotz-Wagenblatt, A., and Ast, G. (2007). Comparative analysis of transposed element insertion within human and mouse genomes reveals alu's unique role in shaping the human transcriptome. *Genome Biol*, **8**(6), R127.

Semon, M. and Wolfe, K. H. (2007). Rearrangement rate following the whole genome duplication in teleosts. *Mol Biol Evol*, **11**, 11.

Seoighe, C. (2003). Turning the clock back on ancient genome duplication. *Curr Opin Genet Dev*, **13**(6), 636–43.

Seoighe, C. and Gehring, C. (2004). Genome duplication led to highly selective expansion of the arabidopsis thaliana proteome. *Trends Genet*, **20**(10), 461–4.

Seoighe, C., Gehring, C., and Hurst, L. D. (2005). Gametophytic selection in arabidopsis thaliana supports the selective model of intron length reduction. *PLoS Genet*, **1**(2), e13.

Shakhnovich, B. E. and Koonin, E. V. (2006). Origins and impact of constraints in evolution of gene families. *Genome Res*, **16**(12), 1529–36.

She, X., Jiang, Z., Clark, R. A., Liu, G., Cheng, Z., Tuzun, E., Church, D. M., Sutton, G., Halpern, A. L., and Eichler, E. E. (2004). Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, **431**(7011), 927–30.

Shi, P., Bakewell, M. A., and Zhang, J. (2006). Did brain-specific genes evolve faster in humans than in chimpanzees? *Trends Genet*, **13**, 13.

Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of arabidopsis thaliana. *Proc Natl Acad Sci U S A*, **99**(21), 13627–32.

Sistrunk, M. L., Antosiewicz, D. M., Purugganan, M. M., and Braam, J. (1994). Arabidopsis tch3 encodes a novel ca2+ binding protein and shows environmentally induced and tissue-specific regulation. *Plant Cell*, **6**(11), 1553–65.

Skrabanek, L. and Wolfe, K. H. (1998). Eukaryote genome duplication - where's the evidence? *Curr Opin Genet Dev*, **8**(6), 694–700.

Snel, B., Bork, P., and Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nat Genet*, **21**(1), 108–10.

Stamm, S. (2002). Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Hum Mol Genet*, **11**(20), 2409–16.

Stamm, S. (2008). Regulation of alternative splicing by reversible protein phosphorylation. *J Biol Chem*, **283**(3), 1223–7.

Storchova, Z., Breneman, A., Cande, J., Dunn, J., Burbank, K., O'Toole, E., and Pellman, D. (2006). Genome-wide genetic analysis of polyploidy in yeast. *Nature*, **443**(7111), 541–7.

Su, Z., Wang, J., Yu, J., Huang, X., and Gu, X. (2006). Evolution of alternative splicing after gene duplication. *Genome Res*, **16**(2), 182–9.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, **278**(5338), 631–7.

Taudien, S., Ebersberger, I., Glockner, G., and Platzer, M. (2006). Should the draft chimpanzee sequence be finished? *Trends Genet*, **22**(3), 122–5.

Taylor, J. S., Van de Peer, Y., and Meyer, A. (2001). Genome duplication, divergent resolution and speciation. *Trends Genet*, **17**(6), 299–301.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (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**(22), 4673–80.

Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A., and Van de Peer, Y. (2004). Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci U S A*, **101**(6), 1638–43.

Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The origins of genomic duplications in arabidopsis. *Science*, **290**(5499), 2114–7.

Waddell, P. J., Kishino, H., and Ota, R. (2001). A phylogenetic foundation for comparative mammalian genomics. *Genome Inform Ser Workshop Genome Inform*, **12**, 141–54.

Wang, X., Grus, W. E., and Zhang, J. (2006). Gene losses during human origins. *PLoS Biol*, **4**(3), e52.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, **23**(13), i549–58.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915), 520–62.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**(4356), 737–8.

Wendel, J. F., Cronn, R. C., Alvarez, I., Liu, B., Small, R. L., and Senchina, D. S. (2002). Intron size and genome size in plants. *Mol Biol Evol*, **19**(12), 2346–52.

Wong, G. K., Passey, D. A., Huang, Y., Yang, Z., and Yu, J. (2000). Is "junk" dna mostly intron dna? *Genome Res*, **10**(11), 1672–8.

Wooding, S. and Jorde, L. B. (2006). Duplication and divergence in humans and chimpanzees. *Bioessays*, **28**(4), 335–8.

Wu, X. W., Lee, C. C., Muzny, D. M., and Caskey, C. T. (1989). Urate oxidase: primary structure and evolutionary implications. *Proc Natl Acad Sci U S A*, **86**(23), 9412–6.

Xing, Y. and Lee, C. (2006). Alternative splicing and rna selection pressure–evolutionary consequences for eukaryotic genomes. *Nat Rev Genet*, **7**(7), 499–509.

Yanai, I., Wolf, Y. I., and Koonin, E. V. (2002). Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol*, **3**(5), research0024.

Yenugu, S., Hamil, K. G., French, F. S., and Hall, S. H. (2006). Antimicrobial actions of human and macaque sperm associated antigen (spag) 11 isoforms: influence of the n-terminal peptide. *Mol Cell Biochem*, **284**(1-2), 25–37.

Yu, W. P., Brenner, S., and Venkatesh, B. (2003). Duplication, degeneration and subfunctionalization of the nested synapsin-timp genes in fugu. *Trends Genet*, **19**(4), 180–3.

Zhang, J., Dean, A. M., Brunet, F., and Long, M. (2004). Evolving protein functional diversity in new genes of drosophila. *Proc Natl Acad Sci U S A*, **101**(46), 16246–50.

Zhang, Z., Sun, H., Zhang, Y., Zhao, Y., Shi, B., Sun, S., Lu, H., Bu, D., Ling, L., and Chen, R. (2006). Genome-wide analysis of mammalian dna segment fusion/fission. *J Theor Biol*, **240**(2), 200–8.

Zietkiewicz, E., Richer, C., and Labuda, D. (1999). Phylogenetic affinities of tarsier in the context of primate alu repeats. *Mol Phylogenet Evol*, **11**(1), 77–83.

Zody, M. C., Garber, M., Adams, D. J., Sharpe, T., Harrow, J., Lupski, J. R., Nicholson, C., Searle, S. M., Wilming, L., Young, S. K., Abouelleil, A., Allen, N. R., Bi, W., Bloom, T., Borowsky, M. L., Bugalter, B. E., Butler, J., Chang, J. L., Chen, C. K., Cook, A., Corum, B., Cuomo, C. A., de Jong, P. J., DeCaprio, D., Dewar, K., FitzGerald, M., Gilbert, J., Gibson, R., Gnerre, S., Goldstein, S., Grafham, D. V., Grocock, R., Hafez, N., Hagopian, D. S., Hart, E., Norman, C. H., Humphray, S., Jaffe, D. B., Jones, M., Kamal, M., Khodiyar, V. K., LaButti, K., Laird, G., Lehoczky, J., Liu, X., Lokyitsang, T., Loveland, J., Lui, A., Macdonald, P., Major, J. E., Matthews, L., Mauceli, E., McCarroll, S. A., Mihalev, A. H., Mudge, J., Nguyen, C., Nicol, R., O'Leary, S. B., Osoegawa, K., Schwartz, D. C., Shaw-Smith, C., Stankiewicz, P., Steward, C., Swarbreck, D., Venkataraman, V., Whittaker, C. A., Yang, X., Zimmer, A. R., Bradley, A., Hubbard, T., Birren, B. W., Rogers, J., Lander, E. S., and Nusbaum, C. (2006). Dna sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature*, **440**(7087), 1045–9.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theor Biol*, **8**(2), 357–66.