MUTATION, DUPLICATION, AND SELECTION IN MAMMALIAN GENOMES

Steven Laurie

TESI DOCTORAL UPF / 2013

DIRECTORA DE LA TESI

Dra. Maria del Mar Albà
Evolutionary Genomics Group,
Research Programme on Biomedical Informatics (GRIB),
IMIM (Hospital del Mar Medical Research Institute) and
Universitat Pompeu Fabra
Barcelona

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES

This thesis is dedicated to the three woman who have helped define the man I am,

my mother Margaret, mi amor Natalia, y mi pequeña Kayla.

Acknowledgements

There are naïve questions, tedious questions, ill-phrased questions, questions put after inadequate self-criticism. But every question is a cry to understand the world. There is no such thing as a dumb question.

Carl Sagan,
The demon-haunted world

This thesis marks the end of a long and winding path. The journey started, as every child's should, alongside my mother, watching David Attenborough's *Life on Earth*, and then with my father, watching Carl Sagan's *Cosmos*. These wonderful television series sparked my interest in the mysteries of life and the universe, and made me want to learn more about, and better understand, how the world works, a feeling that remains with me to this day.

Probably because I found biology easier to understand, and also due to a love of animals, I read Zoology at the University of Glasgow. It was during this time, including a study year abroad at UCLA, that I began to really understand the full implication of Darwin's simple idea. At about the same time molecular genetics was becoming high profile as a result of the Human Genome Project, and I realised that I wanted to move into this exciting new field. Hence I did a Masters by Research in Molecular Genetics at the University of Manchester, which I greatly enjoyed, and subsequently enrolled in the PhD program at Oxford, aiming to study genetics of malaria. However, for a number of reasons this didn't work out, and I became unsure if scientific research was really for me, and began to look into other career avenues.

However, a few years ago, my then gorgeous girlfriend, now wonderful wife, Natalia, convinced me to apply for a place on the Masters in Bioinformatics at UPF, as a step towards returning to research. I found that I was excited to be

back in the academic environment, and from there I came to join Mar Albà's group, and with her support, decided to give a PhD another go.

Firstly therefore, my thanks must go to Mar for taking the chance on a long-in-the-tooth graduate student, for helping write the thesis proposal that resulted in an offer of financial support from the *Generalitat de Catalunya*, and for your guidance, support, and patience throughout the project. I must also thank the other principal investigators who have had an influence on determining the path of the project, Arcadi Navarro, Cedric Notredame, Roderic Guigó, and Toni Galbadón.

Next I must thank my friends and colleagues from the Evolutionary Genomics group with whom I have collaborated during my time here. Maggie, who helped me settle into the lab, and with great patience explained to me how to get data out of Ensembl, how CodeML works, and how to interpret our findings from a biologically meaningful perspective. Núria, who has literally been at my side every step of the way, and would let me distract her with whatever crazy idea was passing through my head at the time. Cinta, who was around too briefly, but was an excellent companion and tireless collaborator during the second part of this thesis. Nico, who taught me many useful programming tricks, and the beauty (in his hands at least) of PHP. José Luis, whose adept Python skills turned an idea into a useful tool in a matter of weeks, and Alice, for always being Alice and a wonderful unofficial *padrina* to Kayla, bringing her a new present every time they meet. I must also thank all of the above for their patience with my poor Castellano, and my non-existent Catalan, and for making such excellent efforts to communicate with me in my mother tongue, rather than their own.

Science is of course a collaborative pursuit, and there are many other individuals who have helped me reach my goal of becoming a Doctor. These include the friends I have made along the way, not least of which are Ignasi Buch, Barbara Montserrat, and Pau Rué, whom I met on the Masters, and each of which has helped with problems along the way, as well as being good friends. Marta Dies falls into this category also, though we met after the masters. I must also thank all the administrative staff; the secretaries from the GRIB, Carina, Chus, and Martina, and their equivalents from IMIM and UPF, who have helped with many different things over the years. In addition, thank you to our excellent Systems Administrators, Alfons and Miguel, for their friendly and efficient resolutions of bugs that have arisen from time to time – real diamonds in their field. After this, there are too many friends and colleagues to mention by name, but thank you in particular to all the people who have attended, and even listened to, my various presentations, and provided valuable feedback. Similarly, thanks to everybody with whom I have shared lunch or coffee, be they friends, or friend's friends, and who make the PRBB such a wonderful environment to work in. A special mention also for the scientists from Darwin's FC, without whom my Saturday afternoons will never be the same, but we all have to hang up our boots one day.

Finally, I must thank my nearest and dearest. My mother Margaret, for all her unconditional support, always. My girlfriend, who during the course of my PhD became my wife, and then the mother of our daughter, Natalia, who has shown extreme tolerance and understanding, particularly during the final, never-ending year of the project - thank you for never giving up on me. My new family, and in particular Jorge & Cristina, who have helped and supported myself and

Natalia in numerous ways during the course of my studies. Last of all, but certainly not least, my daughter Kayla, who made the final year a little more challenging still, but did so without intention, and never failed to bring a smile to her father's face when he returned home, no matter how tired he was, or how tough things got.

If I have forgotten anyone, please forgive me and do not take offence. It is surely through a lack in fully functioning neurons, and not because your help was not valuable and appreciated.

Thank you, each, and every one of you.

Steven Laurie
Barcelona, 10th of June, 2013

ABSTRACT

This thesis comprises comparative genomics analyses primarily focussing on the evolution of mammalian proteins. We concentrate on three species of direct relevance as model organisms, for which high quality genome sequences are available, and human. Having previously investigated protein evolution in terms of substitution rates, here we explored less well studied insertions and deletions (indels). We show that indel and substitution frequencies are correlated at the level of protein sequence, and that indels, and in particular insertions, are elevated in regions of low-complexity and repetitive sequence. Furthermore we observe that selection acts more strongly against the incorporation of insertions than deletions in coding sequence. We also look examine in detail the process of evolution following gene duplication in rodents. We show that in general there is a marked increase in evolutionary rate following duplication, which is restricted to the new copy. We find evidence that this increase is sometimes driven by positive selection, and often accompanied by changes in tissue expression profile. These results lead support to the role of neofuntionalisation following gene duplication.

RESUM

Aquesta tesi consta d'anàlisis de genòmica comparada centrades principalment en l'evolució de les proteïnes de mamífers. Les anàlisis se centren en humans i en tres espècies de gran rellevància com a organismes model, per les quals les seqüències genòmiques són d'alta qualitat. Després d'haver investigat prèviament l'evolució de proteïnes considerant les taxes de substitució, en aquesta tesi hem explorat les insercions i delecions (indels), menys estudiades. Demostrem que existeix una correlació entre la freqüència d'indels i substitucions en la seqüència proteica, i que els indels, i en particular, les insercions, són habituals en les regions de baixa complexitat i seqüències repetitives. A més, observem que la selecció actua més fortament en contra de la incorporació d'insercions que de delecions en la seqüència codificant. D'altra banda, també pretenem analitzar detalladament el procés evolutiu després d'una duplicació gènica en rosegadors. Demostrem que, en general, hi ha un marcat augment en la taxa d'evolució després de la duplicació, que es limita a la nova còpia. I trobem evidències que aquest augment és, de vegades, impulsat per la selecció positiva, i, sovint acompanyada de canvis en el perfil d'expressió de teixits. Aquests resultats recolzen el procés de neofuncionalització després de la duplicació gènica.

PREFACE

Comparative genomics began to come into its own as a field with the advent of relatively cheap and easy DNA sequencing technologies in the 1990s. However, throughout that decade it was limited to relatively small-scale studies, at most amounting to comparisons of tens of genes or proteins across a handful of species. Following the advent of capillary electrophoresis in the late 1990s, and the publishing of the draft human genome in 2001, the number of genomes available for analysis has grown exponentially, and as a result more than 1,000 papers per year on comparative genomics have been published over the last decade, marking it as a truly 21st Century science.

The growth in the field is not solely a result of the number of completely sequenced genomes that are now available. It has been accompanied by many other necessary technological advances, particularly within the field of bioinformatics which has shown even greater growth over the same period. As a result many new tools have been developed for the management and analysis of the huge datasets that sequencing projects have generated, together with large freely-accessible repositories for storing the data generated, that are shared by the international scientific community.

An essential first step in comparative genomics analyses is the generation of multiple sequence alignments, for which a number of useful algorithms have been developed. However, these algorithms do not produce identical results, and thus the first thing I had to do was to to undertake an in-depth analysis to establish which algorithm would be most suitable for the mammalian protein dataset we wished to analyse. I found that a relatively new algorithm, PRANK,

performed much better than competing algorithms in this respect, and was particularly accurate in indel identification, an aspect where other aligners tend to have problems.

As a result of evolution, biological sequences can change in two basic manners; firstly the letters can be substituted for other letters, and secondly, they can change in length, becoming longer or shorter as a result of insertion or deletion of sequence. The vast majority of studies to date have concentrated on substitutions, with relatively few investigating the impact of insertions or deletions (indels). Thus, for the first part of this thesis we decided to investigate the frequency of occurrence of indels in mammalian proteins, and to look for associations with substitutions. The availability of a range of high-quality genomes from closely related mammals allowed us to examine these evolutionary processes in unprecedented detail, and not only did we consider events in the extant species, but we were also able to make deductions about historic events in ancestral branches of the evolutionary tree.

In the second part of the thesis we decided to look in more detail at the first steps in the evolution of a gene newly formed as a result of gene duplication. A number of different models have been proposed that describe the manner in which new genes may evolve following duplication. In particular, there are two major modes by which new genes may adopt a function; subfunctionalisation, which is the partitioning of multiple prior functions between duplicates, and neofunctionalisation, which is the gain of a new function in one of the two duplicates. The degree to which each of these processes contribute to evolution following gene duplication is currently a subject of much debate. Here we have used a set of young gene duplicates from rodents to investigate the tempo and

mode of evolution following duplication, and to gain further insight into this question.

The recent expansion in sequence data available for analysis, from species throughout the tree of life, is finally providing us with the information necessary to test the predictions of theories pertaining to the tempo and mode of evolution that have been proposed over the decades since the development of the *Modern Synthesis* three-quarters of a century ago. The results presented in this thesis represent a small advancement in our knowledge of evolutionary processes within mammalian genomes. As more genomes are sequenced to higher quality, there will be scope for extending these analyses to establish whether the findings here represent general truths, or whether variation between evolutionary lineages tends to be the norm.

Barcelona, June 2013

CONTENTS

1 INTRODUCTION

1.1 Biological Sequences

Biological sequences are ubiquitously represented as strings of letters, with each unique monomeric residue given a particular letter. Thus nucleic acid sequences are composed of strings consisting of the five letters A, C, G, T, and U, representing the 5 common nucleotides (nt), and protein sequences of strings consisting of 20 letters representing the 20 standard amino-acids. As a result of spontaneous mutation, these sequences may change in one of three basic manners; the residue at a particular position in the chain may change to another residue, termed substitution, or one or more residues may be added or removed from the chain, termed insertion or deletion, respectively. Other types of mutation are possible at the level of the genome, including inversion and translocation, which do not affect the sequence of the string itself, but rather its relative position within the genome. However, for the purpose of this thesis I will be considering only point substitutions, and short (<30nt in length) insertions and deletions.

When a mutation occurs in a particular individual, or better stated in the germline that leads to that individual, the change may be beneficial or detrimental, often referred to as advantageous or deleterious respectively, or neutral. A neutral mutation, by definition, will not be subject to natural selection; whether it increases in frequency or disappears completely from a population will be determined by chance alone, depending solely on the effective population size when it arises, a phenomenon known as genetic drift.

If sufficient generations pass, even neutral mutations may become fixed[1]. However, if a mutation is beneficial or detrimental to the organism, then the probability of it being passed on to the next generation is significantly increased or decreased respectively, relative to that of genetic drift alone, in line with the magnitude of its effect on fitness[2]. In the worst case scenario a mutation is lethal, and the affected individual doesn't survive to reproductive age, thus immediately purging the mutation from the population. However, even in the best case scenario, where a mutation is highly advantageous to the individual concerned, chance may nevertheless play a role since the individual may die as a result of accident or predation, in spite of being particularly fit, before it has had the opportunity to reproduce and pass on the beneficial mutation. Thus the vast majority of mutations that arise never make it to fixation.

1.1.1 Neutral sequence

When considering sequence evolution, it is important to keep in mind that the number of observed fixed mutations is much lower than the number of mutation events that have actually occurred. This is of particular importance with respect to protein sequences, in which, depending upon the function of the protein, mutations that result in a non-synonymous[3] substitution may be lethal, and thus will never be observed. However, due to the degenerate nature of the genetic code, many nucleotide substitutions are neutral at the coding level because the affected codon still codes for the same amino-acid – such changes are known as

1 A mutation or allele becomes *fixed* when it reaches a frequency of 100% in the population under consideration.

2 Fitness is measured in terms of the number of offspring produced by an individual that will survive to reach reproductive age, relative to that of the population as a whole.

3 A *non-synonymous* mutation is a substitution in coding DNA that results in a concomitant change in amino acid sequence at the level of the protein.

synonymous substitutions. Nevertheless, not all synonymous substitutions will be completely neutral, as the underlying DNA sequence may itself be involved in gene regulation in some manner; a simple example being mutations that occur in the terminal three positions of an exon thus affecting consensus splice-site recognition (see Chamary *et al*, 2006 for a review of this topic). In order to identify background rates of mutation fixation we would like to be able to identify sequence that is absolutely neutral with respect to selection. However, as sequence that is truly neutral will be evolving very rapidly, we then face the problem of being able to confidently identify homologous sequences between the different genomes under investigation.

Different methodologies have been proposed to attempt to address this issue, each of which has associated problems: use of homologous introns, e.g. in the case of genes with only one intron (Kuo & Ochman, 2009); use of four-fold degenerate sites within coding sequence (Hardison *et al*, 2003); use of ancestral repeat sequences (Imamura *et al*, 2009, Oldmeadow *et al*, 2010,). While the first two techniques address the homology issue fairly well, they fall down a little on their assumption of neutrality and available sample size. Introns are known to include regulatory sequences, including the branch-site motif which is involved in splice-site recognition. Furthermore, it has been shown that the first intron of multi-exon genes tends to be better conserved than other introns, suggesting that they contain other regulatory features too (Gaffney & Keightley, 2006). While four-fold degenerate and other non-synonymous sites may in general be neutral, those found at exon termini will clearly not be for reasons mentioned above, and it is becoming clear that codon-usage bias is reasonably prevalent, suggesting that not all synonymous codons are equivalent within a

genome and thus these sites cannot be considered truly neutral either (Novoa & Ribas de Pouplana, 2012).

Ancestral repeats are a class of common genomic repeat that are the remnants of transposable elements and thus known to have originated from a common sequence (Waterston *et al*, 2002; Hardison *et al*, 2003). They are considered ancestral if they are found in all of the species under investigation but are no longer actively replicating, indicating that they became fixed in position prior to lineage segregation. As a result of the recent burst in whole genome sequencing projects, a very large number of such families have been identified, particularly within mammals (Jurka *et al*, 2005). Since the repeats are ancestral they will be expected to be found in corresponding syntenic regions across the species of interest. Hence they will often be identifiable even when they have diverged substantially, and since they represent transposon remnants, in general they are not expected to play any functional role, and thus will be invisible to selection. However, identification of homologous ancestral repeats does require that there is some degree of sequence conservation. For example, Waterston *et al* (2002) found that RepeatMasker (Smit, 1999) began to have significant problems in detecting ancestral repeats once sequence divergence was greater than 37%. Therefore even mutation rate estimates from ancestral repeat regions will provide only a lower bound for the actual genomic rate, which is also known to exhibit regional variation across the genome (Hardison *et al*, 2003, Ponting & Hardison, 2011). It should also be noted that there are some exceptional cases where it appears that ancestral repeat sequences have been co-opted to provide some sort of regulatory role, as in the case of the MER121 subfamily, the sequence of which has been strongly conserved across mammals and is

therefore believed to perform a cis-regulatory or structural role (Kamal *et al*, 2006).

Here I chose to use ancestral repeats as the preferred method for estimating background rates of mutation fixation due to the large number of sequences available, which are spread throughout the genomes of interest, thus hopefully counteracting any bias due to localised variation in rates of mutation, selection, and recombination (Ellegren *et al*, 2003; Hardison *et al*, 2003; Hodgkinson & Eyre-Walker 2011).

1.2 Multiple sequence alignment

Multiple sequence alignment (MSA) refers to both the process and the product of aligning three or more biological sequences which are assumed to have some degree of homology. The necessity for MSA tools arose following the development of PCR by Kary Mullis in the 1980s (Mullis *et al*, 1986), and the subsequent exponential increase in the identification of biological sequences that continues to this day. MSA has since become the crucial first step in any comparative analysis of molecular sequences, prior to subsequent inference of biologically relevant information such as phylogenetic relationships, molecular function, evidence of natural selection *etc*.

The objective of MSA algorithms is to identify homologous residues in the sequences being analysed, and align them into columns. If successful, each column in the output will reflect the historic path of evolution at that point in the sequence in the species under consideration, thus facilitating identification

of potentially interesting regions of sequence conservation or divergence for further investigation. For regions of sequence which are predicted to have no homology with any other sequence in the alignment (*i.e.* where insertions and/or deletions are assumed to have occurred), gaps, typically represented by hyphens, are inserted in the corresponding column(s) (Figure 1.1).

```
ENSP00000318196     MSEAYFRVESGALGPEENFLSLDDILMSHEKLPVRTETAMPRLGAFFLERSAGAETDNAV
ENSMMUP00000011948  MSEAYFRVESGALGPEENFLSLDDILMSHEKLPVRTETAMPRLGAFFLERSAGAETDNAV
ENSMUSP00000034094  MSEAYFPVESGALGPEENFLSLDDILMSQEKLPVRVETPMPRLGAFFLERGAGSEPDHPL
ENSRNOP00000015795  MSEAYFPVESGALGPEENFLSLDDIVMSQEKLPVRVETPMPRLGAFFLERGAGAEADHPL
ENSBTAP00000012322  MSEAYFRVESGALGPEENFLSLDDILMSHEKLSVRTEIPMPRLGAFFLDRSGGAETDNAI
                    ****** ******************:**:***.**.* .*********:*..*:*.*:.:

ENSP00000318196     PQGSKLELPLWLAKGLFDNKRRILSVELPKIYQEGWRTVFSADPNVVDLHKMGPHFYGFG
ENSMMUP00000011948  PQGSKLELPLWLAKGLFDNKRRILSVELPKIYQEGWRTVFSADANVVDLHKMGPHFYGFG
ENSMUSP00000034094  PQGTKLELPLWLAKGLFDHKRRILSVELPKMYQEGWRTVFSADANVVDLHKMGPHFYGFG
ENSRNOP00000015795  PQGTKLELPLWLAKGLFDNKRRILSVELPKMYQEGWRTVFSADANVVDLHKMGPHFYGFG
ENSBTAP00000012322  PEGTKLELPLWLAKGLFDNKRRILSVELPKIYQEGWRTVFSADANVVDLHKMGPHFYGFG
                    *:*:**************:***********:************.****************

ENSP00000318196     SQLLHFDSPENADISQSLLQ--TFIGRFRRIMDSSQNAYNEDTSALVARLDEMERGLFQT
ENSMMUP00000011948  SQLLHFDSPENADISQSLLQAITFIGRFRRIMDSSQNAYNEDTSALVARLDEMERGLFQT
ENSMUSP00000034094  SQLLHFDSPENADISQSLLK--TFIGRFRRIMDSSQNSYNEDTSALVARLDETERGLFQI
ENSRNOP00000015795  SQLLHFDSPENSDISQSLLQ--TFIGRFRRIMDSSQNSYNEDTSALVARLDETERGLFQI
ENSBTAP00000012322  SQLLHFDSPENADISHSLLQ--TFVGRFRRIMDSSQNAYNEDTSALVARLDEMERGLFQT
                    ***********:***:***:  **:************:************** ******

ENSP00000318196     GQKGLNDFQCWEKGQASQITASNLVQNYKKRKFTDMED
ENSMMUP00000011948  GQKGLNDFQCWEKGQASQITASNLVQNYKKRKFTDMED
ENSMUSP00000034094  GQRSLNDFQSWEKGQASQITASSLVQNYKKRKFTNMED
ENSRNOP00000015795  GQKGLNDFQSWEKGQASQITASSLVQNYKKRKFTNLED
ENSBTAP00000012322  GQKGLNDFQCWEKGQASQLTASNLVQNYAKRKFTDMED
                    **:.*****.********:***.***** ***** ::**
```

Figure 1.1. Typical MSA output (aln format). Multiple alignment of GINS3 orthologs from human, macaque, mouse, rat and cow. It is clear from the consensus line that the protein sequence has been highly conserved across these species since 84% of the columns are identical (asterisks). However, there is a very clear insertion of two amino acids in the macaque ortholog in position 141-142, but see also Figure 3.1.

MSA algorithms attempt to maximise a score which is affected by two key parameters:

1) A substitution matrix which provides a score based on the probability that two residues should be aligned. The most commonly used matrices for scoring amino acid MSAs are those of the BLOSUM family (Figure 1.2), which were determined by empirical observation of pairs of known orthologous sequences of different degrees of similarity across different species (Henikoff & Henikoff, 1992). Matrices for aligning DNA sequences are much more straightforward, generally assuming any change to be equally likely, or incorporating simple weighting for the difference in the probability of transition or transversion substitutions.

2) A gap penalty that defines the cost for placement of a gap in cases where there is deemed to be no homologous residue in a particular sequence. Gap penalties may be further refined in terms of having separate gap-opening, and gap-extension penalties, and by dynamically varying these penalties.

By summing the appropriate scores across an alignment, the MSA algorithm determines the optimal alignment, and this is typically output to the user in the form of columns of aligned sequence together with a line indicating degree of conservation (Figure 1.1).

However, in practice the process is not as straightforward as outlined above. Whereas the process of aligning a pair of sequences, incorporating a scoring function and an appropriate gap-penalty, can be optimised to guarantee the best

possible alignment using the Needleman and Wunsch (1970) implementation of a dynamic programming algorithm, application of dynamic programming to multiple sequences of significant length is extremely computationally expensive. Thus all current MSA algorithm implementations use various heuristics to simplify the process, by far the most commonly applied of which is progressive alignment (Feng & Doolittle, 1987).

	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
A	**5**	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	**6**	-1	-2	-4	1	-1	-3	0	-3	-3	2	-2	-4	-2	-1	-1	-4	-3	-3
N	-2	-1	**6**	1	-3	0	-1	-1	0	-4	-4	0	-3	-4	-3	0	0	-4	-3	-4
D	-2	-2	1	**6**	-4	-1	1	-2	-2	-4	-5	-1	-4	-4	-2	-1	-1	-6	-4	-4
C	-1	-4	-3	-4	**9**	-4	-5	-4	-4	-2	-2	-4	-2	-3	-4	-2	-1	-3	-3	-1
Q	-1	1	0	-1	-4	**6**	2	-2	1	-3	-3	1	0	-4	-2	0	-1	-3	-2	-3
E	-1	-1	-1	1	-5	2	**6**	-3	0	-4	-4	1	-2	-4	-2	0	-1	-4	-3	-3
G	0	-3	-1	-2	-4	-2	-3	**6**	-3	-5	-4	-2	-4	-4	-3	-1	-2	-4	-4	-4
H	-2	0	0	-2	-4	1	0	-3	**8**	-4	-3	-1	-2	-2	-3	-1	-2	-3	2	-4
I	-2	-3	-4	-4	-2	-3	-4	-5	-4	**5**	1	-3	1	-1	-4	-3	-1	-3	-2	3
L	-2	-3	-4	-5	-2	-3	-4	-4	-3	1	**4**	-3	2	0	-3	-3	-2	-2	-2	1
K	-1	2	0	-1	-4	1	1	-2	-1	-3	-3	**5**	-2	-4	-1	-1	-1	-4	-3	-3
M	-1	-2	-3	-4	-2	0	-2	-4	-2	1	2	-2	**6**	0	-3	-2	-1	-2	-2	1
F	-3	-4	-4	-4	-3	-4	-4	-4	-2	-1	0	-4	0	**6**	-4	-3	-2	0	3	-1
P	-1	-2	-3	-2	-4	-2	-2	-3	-3	-4	-3	-1	-3	-4	**8**	-1	-2	-5	-4	-3
S	1	-1	0	-1	-2	0	0	-1	-1	-3	-3	-1	-2	-3	-1	**5**	1	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-2	-1	-1	-2	-2	1	**5**	-4	-2	0
W	-3	-4	-4	-6	-3	-3	-4	-4	-3	-3	-2	-4	-2	0	-5	-4	-4	**11**	2	-3
Y	-2	-3	-3	-4	-3	-2	-3	-4	2	-2	-2	-3	-2	3	-4	-2	-2	2	**7**	-2
V	0	-3	-4	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-2	**4**

Figure 1.2. BLOSUM80 Matrix. The matrix is symmetrical about the diagonal, and provides a log-odds score of the probability of observing a particular amino acid at a particular position in a protein, given the residue observed at the same position in an orthologous protein, and assuming the similarity overall between the orthologs is 80%. The high values on the diagonal indicate that the most likely observation is that the amino-acid will be identical at a particular position between two such orthologs. Furthermore, the matrix shows that tryptophan (W) and cysteine (C) are the best conserved of all amino acids. Interestingly they are also the least frequently observed amino acids in nature. (after Henikoff & Henikoff, 1992)

Progressive alignment involves initial pairwise alignment of all possible pairs of sequences in order to establish a guide tree of relatedness, and subsequent incorporation of sequences into the multiple alignment in order of decreasing relatedness. The first widely-used implementation of progressive alignment in addressing the MSA problem was the development of the CLUSTAL series of algorithms (Higgins & Sharp, 1988), and in particular CLUSTALW (Thompson *et al*, 1994) which remains by far the most widely cited MSA algorithm (Table 1.1) despite performing significantly worse than any of its competitors. Subsequently a number of improvements upon the basic progressive alignment algorithm have been developed, resulting in higher-scoring alignments. These include T-Coffee (Notredame *et al*, 2000), MAFFT (Katoh *et al*, 2002), and MUSCLE (Edgar, 2004). Each of these applies adjustments to the scoring scheme, primarily with regards to gaps, together with iterative realignment of problematic regions to improve the quality of the MSA produced, and depending upon which author you read, each out-performs the others.

MSA Algorithm	Original Paper	Google Scholar	Pubmed[a]	Pubmed 2008-2012
ClustalW	Thompson *et al*, 1994	41,221	8,707	3,742
MUSCLE	Edgar, 2004	5,607	2,553	2,160
T-Coffee	Notredame *et al*, 2000	3,396	1,090	637
MAFFT	Katoh *et al*, 2002	1,585	552	430
PRANK	Löytynoja & Goldman, 2008	188	87	79

Table 1.1. Number of citations for some popular multiple sequence alignment algorithms. Note that ClustalW has been cited more often than all the other programs combined, including during the last five years alone. [a]Total citations up to May 15, 2013

Nevertheless, different alignment algorithms will often produce quite different alignments, particularly when areas of sequence are less well conserved, thus affecting subsequent analyses and deductions based upon the alignments (reviewed in Wong *et al*, 2008). In particular the placement of gaps in the alignment, implying the presence of an insertion or deletion in a particular sequence, is dependent on the algorithm applied and its associated gap penalties (Higgins *et al*, 2005, Golubchik *et al*, 2007). One artefact common to dynamic programming based algorithms is over-alignment, whereby gaps tend to attract and become clustered together, consequently forcing non-homologous portions of sequence to align (Figure 1.3a). The resulting alignments thus inflate estimates of amino-acid substitution rate, underestimate the number of insertions, and incorrectly estimate the number of deletions that have occurred during evolutionary history, relative to the true values.

As a result of this alignment uncertainty, a number of researchers have attempted to implement methods to estimate and improve the reliability of MSA, or to choose between competing alignments through incorporation of additional information (e.g. Landan & Graur, 2007; Muller *et al*, 2010, Penn *et al*, 2010). Such methods generally involve comparing either the sum-of the pairs score (*i.e*. the percentage of correctly aligned residue pairs in the MSA), or the column-score (*i.e.* the percentage of correctly aligned columns in the alignment) (Thompson *et al*, 1999a) against some form of test set or benchmark, a popular choice of which is BALIBASE (Thompson *et al*, 1999b). An alternative approach is simply to remove regions of alignments that are of lower quality/confidence (*i.e.* typically those regions where there are a number of gaps) from subsequent analyses and only focus on the better conserved parts

of the alignment in which we have more confidence (Catresana, 2000). However, such measures will result in loss of potentially useful information and are likely to introduce downstream biases. Furthermore, though MSA algorithms will successfully identify the optimal alignment based upon the scoring parameters used, it is often the case that the alignment that truly reflects evolutionary history will be sub-optimal due to the stochastic nature of sequence evolution. It must also be borne in mind that the specific set of parameters or methodological steps that work best for a particular dataset will not necessarily be those that work best for alternative datasets.

Löytynoja and Goldman (2005, 2008) have developed a novel MSA algorithm, PRANK+F (hereafter simply referred to simply as PRANK), that uses a Hidden Markov Model (Eddy, 2004) together with a user-defined species tree to better determine gap positions during the progressive alignment process (Figure 1.3b). It records the observation of any apparent insertions in order to prevent subsequent alignment of further sequences to the corresponding columns in later stages of the alignment process, thus improving the quality of the alignment, and providing more accurate measures of indel frequencies (Dessimoz & Gil, 2010). While this algorithm is approximately an order of magnitude more computationally expensive than competing MSA algorithms, it has been shown to produce MSAs that provide a better representation of the evolutionary history of the sequences under examination (Fletcher & Yang, 2010, Markova-Raina & Petrov, 2011; Jordan & Goldman 2012).

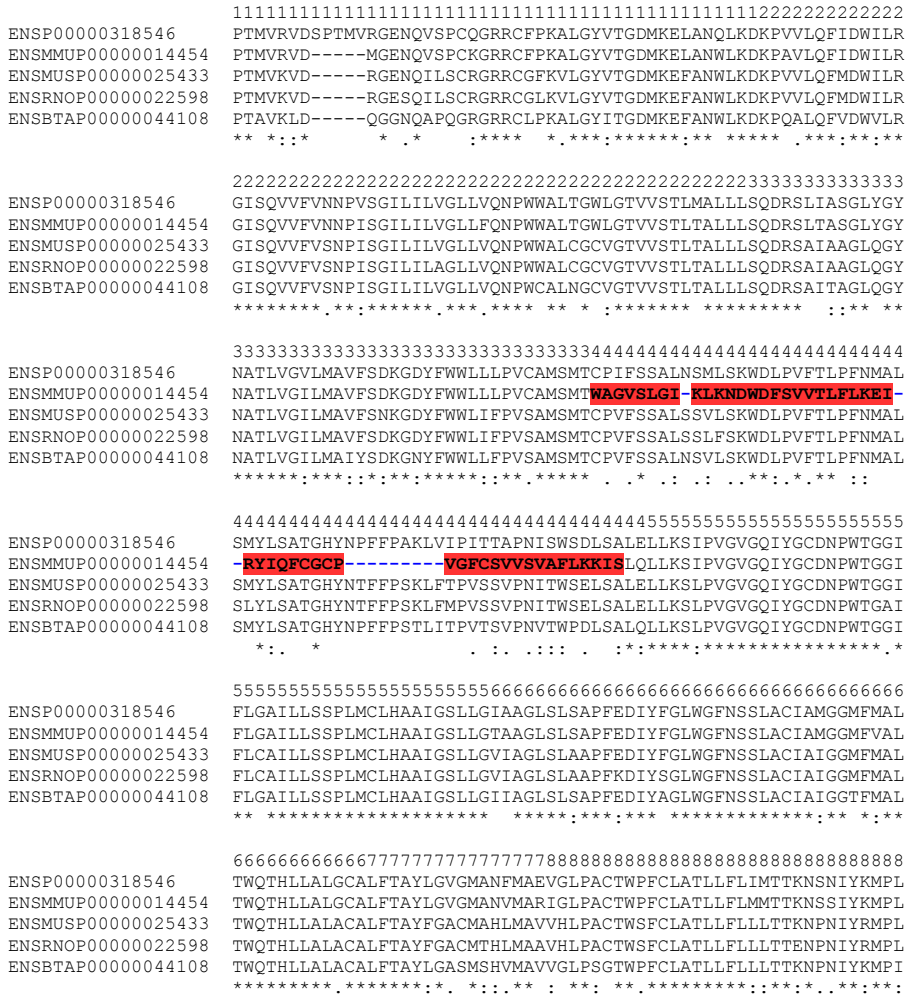


Figure 1.3a MAFFT alignment of the longest isoform of SLC14A1 and its orthologs showing over-alignment of non-homologous sequence (highlighted in red). The numbers above the alignment correspond to the positions of coding exons in the human sequence. The problematic region in macaque is annotated in Ensembl (likely incorrectly) as two short exons that are clearly not homologous to exon 4 of the other species being aligned here. If no further error-checking was performed on this alignment, we would interpret there to have been three short deletions events (in blue), and a large number of substitutions, in the macaque gene, thus resulting in a false-positive observation of positive selection for this protein in macaque. N.B. The first and last blocks of the alignment have been removed for clarity. Compare with PRANK output in Figure 1.3b, overleaf.

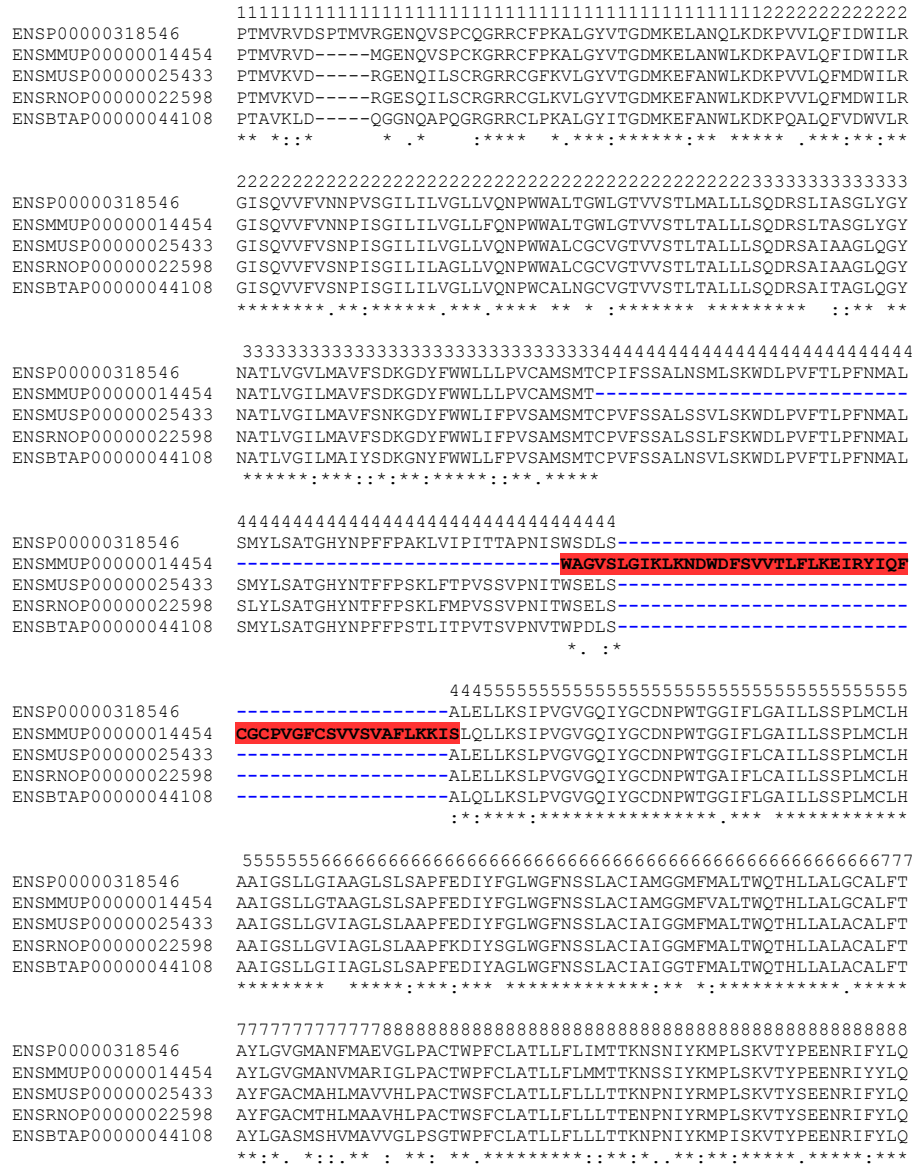


Figure 1.3b PRANK alignment of the longest isoform of SLC14A1 and its orthologs. PRANK correctly identifies the problematic region in macaque as non-homologous and separates it out. However, PRANK does misinterpret chance similarity in the fourth block as homology, due to the tryptophan residue (the most highly conserved amino acid - see Figure 1.2) in all sequences. While the alignment is not perfect, it is much less likely to result in a false-positive observation of positive-selection. Compare with MAFFT output in Figure 1.3a

1.3 Indels

Insertions and deletions (indels) may arise in a number of different manners in the genome including DNA mispairing, non-homologous recombination, non-homologous end-joining, sequence-slippage, and transposition. They range in size from events that affect individual base-pairs, to events affecting megabases of DNA resulting in the duplication or loss of long stretches of chromosome. In general the exact mechanism leading to indel generation is unknown since they are typically inferred *post hoc* through comparison with homologous sequences. However, studies of mammalian genomes have found that some form of tandem matching, be it immediately adjoining or at a short distance from the point of mutation appears to be the norm in the case of insertions, though this appears to be less true of deletions (Taylor *et al*, 2004; Messer & Arndt, 2007; Tanay & Siggia, 2008). One mechanism leading to indel formation that has been well described is that of sequence-slippage of short regions of tandemly repeating sequence during replication (Levinson & Gutman, 1987).

Background sequence context has also been proposed to have an effect on indel occurrence. Tanay & Siggia (2008) reported a bias towards adenine and thymine bases in the immediate vicinity of human indels that was independent of relative local GC content, with adenines tending to precede an indel, and thymines tending to follow the indel. Investigating larger-scale motifs throughout the human genome, Kvikstad *et al* (2007, 2009) suggest that indel rate is affected by local GC content, recombination rate, and replication. They report an association of indels with topoisomerase cleavage sites, implicating recombination, and with DNA polymerase pause sites, implicating replication,

and hence hypothesise that recombination may be more important in generating insertions while replication events more commonly generate deletions. Thus there appears to be a range of mechanisms that are involved in the generation of indels, and that those that result in insertions may be distinct from those that result in deletion.

1.3.1 Frequency of indel occurrence

Early attempts to quantify the frequency of indel events in both coding and non-coding sequences observed an apparent bias towards deletions. de Jong and Ryden (1981) reported a four-fold excess of deletions over insertions in an analysis of 9 families of known homologous proteins, although their final sample size was only 30 events and thus only marginally significant. In a later study of fifty-two purported human and rodent retrotransposed pseudogenes, Graur *et al* (1989) reported a more significant 7-fold bias in humans and 3-fold bias in rodents towards deletions. However, when this study was later extended to include 109 pseudogenes (Ophir and Graur, 1997), the observed biases reduced to 2.9-fold in humans, and 2.6-fold in rodents. Reviewing these and other early studies in invertebrates, led Petrov (2002) to postulate that such a bias towards deletions may have a role in determining equilibrium genome size in the taxa concerned, though this view was subsequently strongly challenged by Gregory (2003, 2004) on the basis that larger-scale events such as retrotransposon family activity are likely to play a more substantial role in determining genome size.

The pre-genomic era studies referred to above undoubtedly suffered from a lack of power in homolog detection and likely used sequence data of limited quality. Nevertheless, a number of studies in the post-genomic era have suggested that the observed bias towards deletion generally holds, though it may be less prominent than first reported. Initial analyses of the mouse (Waterston *et al*, 2002) and rat (Gibbs *et al*, 2004) genomes reported a deletion bias in genomic sequence of approximately 2.5:1 and 3.1:1 respectively, which dropped to 1.1:1 and 1.7:1 for coding sequence (Taylor *et al*, 2004). However, a contemporary study of mouse and rat genomic sequence using different genomic alignments and focussing on indel events of up to just 10bp in length reported biases of only 1.5:1 in mouse and 2:1 in rat (Cooper *et al*, 2004). While Mills *et al* (2006) reported only a 1.1:1 bias in favour of deletions in a genome-wide scan in humans when comparing with the chimpanzee genome, both Kvikstad *et al* (2007) and Tanay and Siggia (2008) found a bias of approximately 1.5:1 across the human genome in comparisons with macaque and chimpanzee sequences. These conflicting findings probably reflect differences in indel identification and filtering methodologies, and perhaps also to differences in the chimpanzee genome builds used. Latterly, Kuo and Ochman (2009) performed a coarse but wide-ranging analysis across Archaea, Bacteria and Eukaryota, the last of which was limited to human, *Drosophila* and *Saccharomyces*, in which they report a deletion bias in all cases, and thus declared that deletion bias was universal across all domains of life.

1.3.2 Correlation between indels and point substitutions

A common finding in many of the aforementioned studies is a correlation between substitution rates and indel occurrence, suggesting that particular areas of the genome may be more prone to mutation and/or more tolerant of mutation. The first to report such a correlation were Gu & Li (1992) in a comparative study of 54 orthologous proteins from human and mouse or rat, using chicken as an outgroup, in which they identified 75 unpolarised indels and found that in general the rodent branch was evolving approximately 50% faster than the human branch, both in terms of amino acid substitutions and in number of accumulated indel events. However, when the initial mouse and rat draft genomes were completed a little over a decade later, it was reported that the rodent lineages have evolved almost three times as fast as the human lineage in terms of substitution and deletion, though only 2.3 times as fast in terms of insertions (Gibbs *et al*, 2004). Interestingly, while there was found to be a reasonable correlation between each of these three classes of mutation, substitutions were found to correlate more strongly with deletions (R^2~0.40), than with insertions (R^2~0.25).

Intriguingly, Tian *et al* (2008) have put forward the hypothesis that indels may be mutagenic towards surrounding sequence. They found that the rate of substitution around indels in a variety of eukaryotic genomes is elevated proximally to indels and suggest that this may be a result of indels being mutagenic during meiosis while segregating in a heterozygous state. If Tian and colleagues' hypothesis is correct, then it may explain many observed genomic

correlations, including the association between indel rates and point mutation, lower mutation rates on sex chromosomes (since they are largely hemizygous), and why organisms with short generation times and larger effective population sizes tend to have higher mutation rates (see Hodgkinson & Eyre-Walker, 2011 for further discussion). However Tóth-Petróczy and Tawfik (2013) observing a similar pattern in a comparative analysis of protein sequence in yeast species, have suggested the reverse relationship i.e. neutral substitutions precede indel occurrence. Curiously, they found no such correlation in non-coding regions and thus further investigation of this matter will be required in order to understand better the order of events.

1.4 Positive Selection

Natural selection, as first presented to the Linnean Society of London in 1858 by Charles Darwin and Alfred Russell Wallace, and subsequently gaining fame through Darwin's defining treatise *On the Origin of Species by Means of Natural Selection* (1859), is the process whereby individuals in a population that are better adapted to their environment as a result of heritable traits will tend to have more offspring than will less well adapted individuals, and thus these favourable traits will increase within the population. Ever since the deciphering of the genetic code in the 1960s and the advent of the field of molecular evolution, evolutionary biologists have been interested in identifying the signature of natural selection in biological sequences. In particular, we have been interested in identifying cases of positive selection *i.e.* where a new mutation arises that increases the fitness of affected individuals, and thus

spreads throughout a population until it eventually becomes fixed, or where selection acts upon standing genetic variation in a population when a particular allele becomes strongly advantageous in a particular environment, pushing the frequency of that allele towards fixation.

In the context of this thesis, where I am mostly considering protein coding sequences across species, by definition I am considering fixed differences between species, though it is likely that a small fraction of the observed differences may be polymorphic in some of the species concerned. However, simple observation of a difference in sequence between species is not evidence of positive selection, as it most likely will have arisen by genetic drift alone (Kimura, 1968; King & Jukes, 1969). Indeed, the majority of observed mutations are believed to be neutral or slightly deleterious (Ohta, 1973), since seriously deleterious mutations can never reach a high frequency in the population and will most likely be rapidly purged, while advantageous mutations are believed to be rare.

Thus in order to be able to detect plausible evidence of positive selection when analysing sequence data alone, we need to be able to show that observed differences at the sequence level are not likely to have occurred solely as a result of drift. One way to do this is to measure if there is an excess of non-synonymous substitutions compared to synonymous substitutions in coding sequences. However, this is not as straightforward as it might appear for a number of reasons: non-synonymous substitutions are about thrice as likely as synonymous substitutions to occur by chance due to the nature of the genetic code, assuming the resultant change is selectively neutral; transition

substitutions (A => G / C => T or *vice versa)* are more commonly observed than transversions (the other eight possible mutations); when two changes are observed in a particular codon, the two appropriate paths need to be weighted distinctly depending on their relative likelihoods; when comparing sequences it is not possible to be certain how many changes have occurred (*i.e.* a C at a particular position in one lineage may be a G in another, but it is possible that it was an A or T between times); as phylogenetic distance increases, saturation of changes becomes a problem.

These issues, and how best to deal with them, kept various investigators busy throughout the final quarter of the 20th century (see Chapter 4 of Li (1997) for a detailed review). The methodology that has subsequently gained most traction among the evolutionary biology community is the implementation of a likelihood ratio test within a maximum likelihood framework, for detecting positive selection at the codon level, as performed by the CodeML program of the PAML package (Yang, 1997; Yang, 2007), though its validity has occasionally been questioned (Nozawa *et al*, 2009; Wolf *et al*, 2009). In particular the branch-site model, initially developed in 2002 (Yang & Nielsen), and later improved in 2005 (Yang *et al*, 2005; Zhang *et al*, 2005) has been widely used in comparative genomics to test for positive selection. This test attempts to take into account all of the problematic variables mentioned above, and tests for selection at the codon level in a particular branch specified by the user. In practice, in analyses utilising a small number of taxa the user will typically test all branches individually and then apply a correction for multiple testing, since there is often no *a priori* hypothesis as to which branch positive selection is expected to have acted upon.

Simulations have shown that the test is generally conservative (Yang & dos Reis, 2011), and thus may miss instances of positive selection, but the authors argue correctly that this is preferable to the generation of a surplus of false positives. Importantly, Yang and dos Reis (2011) and others (e.g. Schneider *et al*, 2009) have shown that the branch-site test is highly sensitive to alignment quality, with poor quality alignments generally resulting in a marked inflation in the number of false-positives due to over-alignment of non-homologous residues. Schneider *et al* (2009) also found that both gene annotation status and raw sequence trace quality greatly affected estimates of the number of genes appearing to have undergone positive selection.

The importance of alignment quality is highlighted by early analyses involving the initial assembly of the chimpanzee genome which suggested that there have been many more instances of positively selected genes in the chimpanzee lineage than in the human lineage since they diverged (Bakewell *et al* 2007, Gibbs *et al*, 2007). This was subsequently shown to reflect poor-quality raw sequence, incorrect homology identification, and alignment ambiguity in the vicinity of indels and breaks in the genomic alignments (Mallick *et al* 2009). Similarly, initial reports suggesting that 11.5% of chimpanzee genes have undergone positive selection since divergence from humans when using MUSCLE for aligning (Vamathevan *et al*, 2008), were reduced to an estimate of just 2.7% when PRANK was used to regenerate the same alignments (Fletcher & Yang, 2010). Given that the majority of vertebrate genomes currently available in public databases are of quality equal to, or less than, that of the initial chimpanzee build, these issues must be carefully borne in mind when undertaking large-scale comparative genomic analyses.

1.5 Gene Duplication

The concept of gene duplication as an evolutionary process dates back at least as far as the modern evolutionary synthesis (Bridges, 1936), and is considered to be the primary process leading to the formation of new genes. Alternative sources of new genes include horizontal gene transfer between species, and *de novo* evolution from non-coding sequence. The former of these is extremely rare in animals (Dunning Hotopp, 2011), while the degree of contribution of the latter remains to be established (Knowles & McLysaght, 2009; Toll-Riera *et al*, 2009; Tautz & Domazet-Lošo, 2011).

It is likely that there are many genes that cannot undergo duplication without having a detrimental effect on the organism, as illustrated by the observation that genes with certain characteristics such as high evolutionary rate (prior to duplication), and higher essentiality (estimated using interaction networks and knockout models) are less likely to be observed in duplicate (Davis & Petrov, 2004; He & Zhang, 2006; Li *et al*, 2006; Conant & Wolfe, 2008). Traditionally gene duplication has been ascribed to one of three mechanisms: whole genome duplication (WGD), which results in polyploidy and has been ubiquitous in the evolution of plants (Rieseberg & Willis, 2007), but is also thought to have occurred in the ancestral lineage leading to the vertebrates (Makalowski, 2001; Dehal & Boore, 2005), and again in teleost fish (Jaillon *et al*, 2004); tandem-duplication, whereby a stretch of chromosome is duplicated, typically as a result of unequal crossing over during recombination, and thus any genes contained therein are duplicated; retroduplication, where an mRNA molecule is reverse-

transcribed and reinserted into the genome through the action of a retrotransposon element, producing a retrogene (Deininger & Batzer, 2002). Recently a fourth mechanism has been proposed, called duplicative transposition or drift duplication, which attempts to explain the observation of relatively distant duplicates (*i.e.* non-tandem) which may have arisen as a result of non-allelic homologous recombination (Hahn, 2009; Ezawa *et al*, 2011). However it remains uncertain whether the majority of such cases occur at the moment of duplication, or whether they result from initial tandem duplication followed latterly by chromosomal rearrangement.

Whole genome duplication will not be considered in detail here since it is not relevant to the theme of this thesis, but the outcomes of tandem duplication and retroduplication will be examined. Throughout this section, except where explicitly expressed otherwise, I will be considering duplication of a complete progenitor gene, referred to as the *parent* gene, resulting in the birth of a new identical *daughter* copy. I will return to the special case of retrogenes, which are not created identical to their progenitors, in Section 1.5.3.

Assuming that gene duplication results in a viable organism, there are two fates that can befall the daughter gene, which are analogous to those described for a novel point mutation in Section 1.1 above. Firstly, it may disappear from the population, restoring the prior *status quo*. Secondly, it may increase in frequency by drift and/or selection and eventually become fixed in the population. During the process of fixation, a gene may become non-functional due to the accumulation of detrimental mutations, resulting in a fixed pseudogene. On the other hand, assuming the duplicate remains functional,

there are considered to be three possible endpoints: it may maintain the same role as the ancestor resulting in increased gene dosage (Konrad *et al*, 2011); it may take on a novel role not found in the progenitor, termed neofunctionalisation (Ohno, 1970; Force *et al*, 1999); it may take on only part of the role of the progenitor, termed subfunctionalisation (Ohno, 1970, Force *et al*, 1999). Figure 1.4 provides a schematic representation of these outcomes.

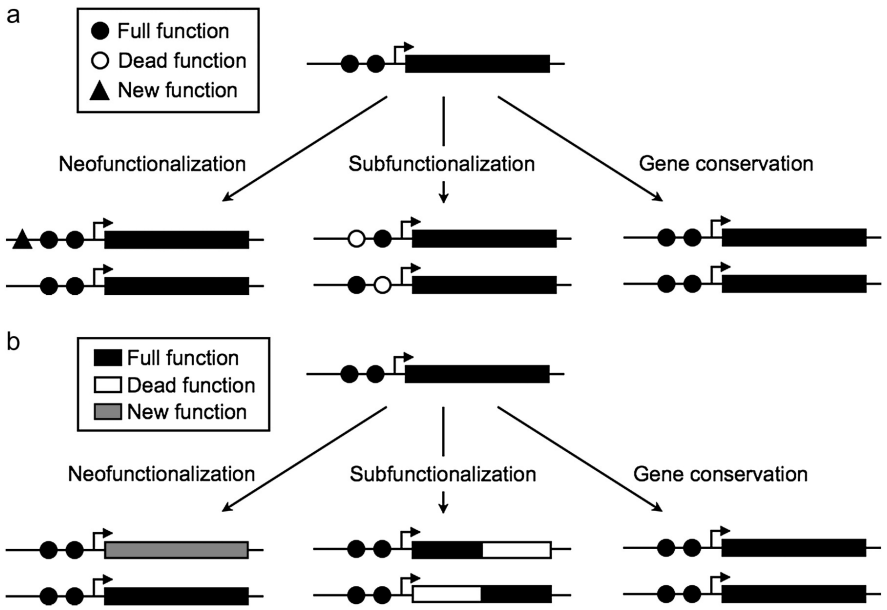


Figure 1.4. Three possible functional outcomes following gene duplication. a. Outcomes following regulatory sequence changes; changes in the promoter may affect gene regulation thus altering temporal and/or spatial patterns of expression (not considered in this thesis). **b.** Outcomes following coding sequence changes; neofunctionalisation may arise from relatively few changes in one of the duplicates, while subfunctionalisation requires changes in both duplicates. Image taken from Hahn (2009).

Thus there are two stages in the evolution of a novel gene duplication: maintenance of the two duplicate copies following the duplication event, and the path to fixation, which may involve the action of selection (Figure 1.5). These processes will be affected by general population genetic parameters, and may overlap in time (Walsh, 2003; Conant & Wolfe, 2008; Innan & Kondrashov, 2010).

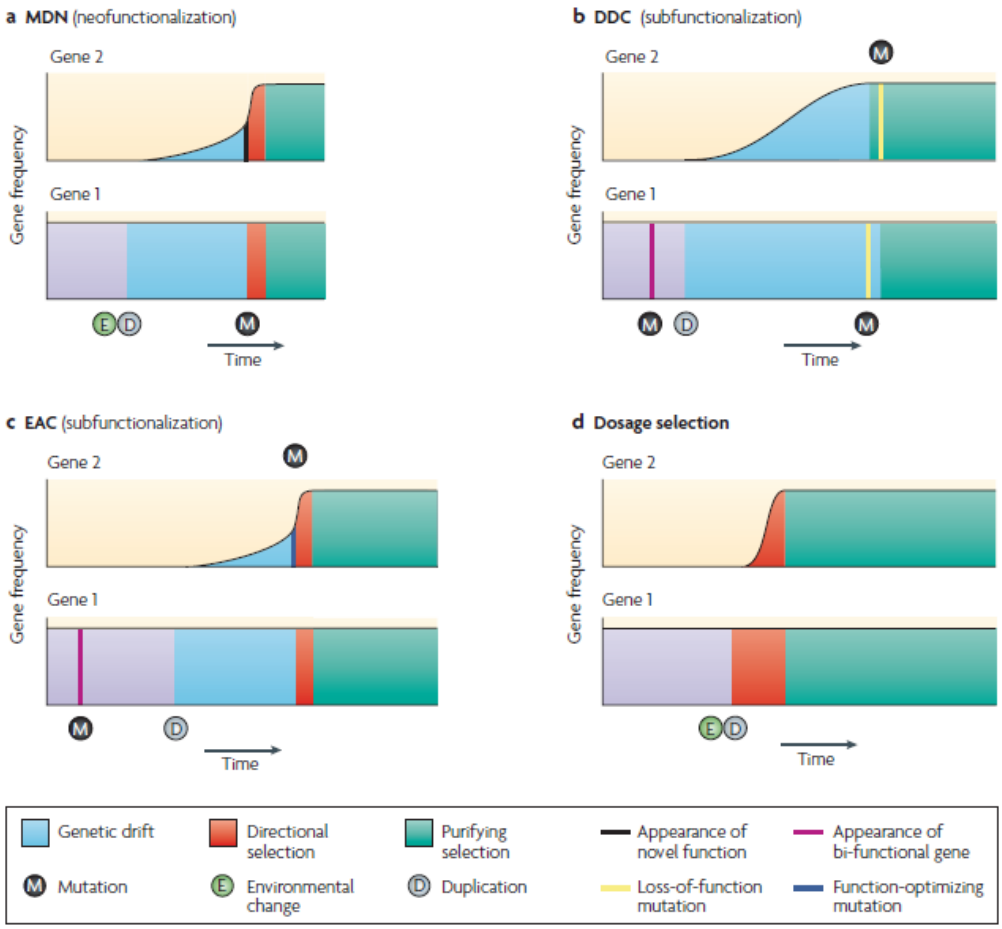


Figure 1.5. Possible mechanisms of fixation following duplication. Each panel shows the frequency of each member of a pair of genes in the population. Gene 1 is fixed prior to duplication, which results in the origin of gene 2. **a.** Here, the frequency of the new copy is initially increasing by drift alone, until a novel mutation resulting in neofunctionalisation arises that is subsequently strongly selected for pulling the gene to fixation. **b.** Force's (1999) DDC model of subfunctionalisation: following fixation of the new copy by drift, complementary degenerative mutations arise in each duplicate independently such that both copies are necessary to maintain ancestral function. **c.** The *Escape from Adaptive Conflict* model (Hughes, 1994) of subfunctionalisation: here mutations in one of the duplicates results in optimisation for one subfunction at the cost of the other(s), and selection results in fixation of this new version. **d.** Dosage selection: environmental change results in increased dosage being beneficial and thus the duplication is strongly selected for and rapidly proceeds to fixation. Note that for cases **b-d**, there is nothing to rule out neofunctionalisation playing a role, subsequent to fixation, in either of the duplicates. Other possible mechanisms exist. Image taken from Conant & Wolfe (2008), text adapted from same.

1.5.1 Maintenance of the initial duplicate

We can be almost certain that the majority of gene duplications that have occurred throughout evolution have left no evidence (Lynch & Conery, 2000). Firstly, it has been shown that not all gene duplications are viable e.g. if a gene in the centre of a tightly regulated pathway or network were to be duplicated, resulting in a doubling of gene product that could not be down-regulated by the cell in any other manner, then this will likely be severely detrimental or even lethal to the organism (Eppig *et al*, 2005; He and Zhang, 2006 ; Liang & Li, 2007). On the other hand, if such an increase in gene product were to provide a large selective advantage to the organism, then it is likely that both the parental and daughter gene will be maintained by positive selection, and thus fixation may occur rapidly with almost no change at the sequence level. A finding that is consistent with this model is that of increased copies of the salivary amylase gene being observed in human populations that consume diets rich in starch (Perry *et al*, 2007). Between these extremes, there are a number of other manners in which an initial duplication may be favourably retained:

1) Having an identical, fully redundant, duplicate copy may mask any deleterious mutations that happen to occur in the parent gene, particularly if they result in loss of function. However models show this effect will be of negligible significance except in very large populations (Clark, 1994).

2) If heterozygosity at the original locus is advantageous (*i.e.* classic genetic overdominance), then gene duplication, if it occurs initially in a

heterozygote individual, may generate a state of *de facto* permanent heterozygosity, and again be strongly advantageous and thus favoured by positive selection. This has been shown to have occurred a number of times independently in the acetylcholinesterase gene of mosquitoes of the genus *Culex* in response to pressure from insecticides (Labbe *et al*, 2007). This may also help to explain the large amount of diversity observed at many loci involved in immune response located within the human major histocompatibility complex region on chromosome six which typically display overdominance (Hughes & Nei, 1989).

3) If a duplicate picks up a chance beneficial allele shortly following duplication then this will clearly favour retention also.

Ohno's classic model (1970) of gene duplication suggests that there is a period of relaxed selection following duplication due to redundancy, in which one copy, typically the novel one, will accumulate degenerate or nonsense mutations and thus pseudogenise. However, very occasionally a beneficial allele may arise during this process, leading to positive selection and neofunctionalisation. One important problem with this model is that since the new copy will be initially at very low frequency within the population, the chance that a beneficial mutation, a very rare event in itself, may happen to occur in the new copy is almost vanishingly small. Thus perhaps we should consider it more likely that the beneficial mutation arises prior to duplication, and subsequently allelic sampling of the beneficial allele occurs following duplication, in a manner similar to case two above. However, if both copies are maintained, more or less unchanged, until the frequency of the daughter copy

reaches a reasonable level within the population, then there will be a higher probability that the beneficial mutation may affect the daughter copy.

1.5.2 Evolution following maintenance

Nonfunctionalisation

Following duplication, if the daughter copy is completely redundant, and thus selectively neutral, it will be free to accumulate mutations rapidly. In the vast majority of cases these mutations will lead to degeneration, and eventual silencing of the copy. Note that if the two copies are exactly equal then it could be that it is the parental locus that eventually becomes nonfunctional, but as chromosomal context usually affects cis-regulation, it is unlikely that duplication often results in copies that are absolutely identical with respect to selection and thus the daughter will typically be slightly less fit, and hence less likely to be maintained. Once the daughter copy has begun to accumulate mutations, we expect there to be continued purifying selection for maintenance of the parent copy, and further relaxation of selection on the daughter as it continues down the path towards pseudogenisation. Currently 39% of all possible protein-coding genes identified in the human genome have been classed as pseudogenes by Ensembl[4] (GRCh37 assembly, gene build April 2011), suggesting that this process has been commonly repeated throughout evolution.

4 Ensembl defines a pseudogene as "*a genomic region that shares an evolutionary history with a protein-coding gene, but has incorporated frame-shifting or stop codon mutations that disrupt an open reading frame*".

Neofunctionalisation

This term was coined by Force *et al* (1999), though the process was first outlined by Ohno (1970). In Ohno's classical setting, following gene duplication and relaxed selection on the daughter copy, by chance a novel beneficial mutation may arise that is subsequently maintained by positive selection and starts the path towards fixation, with the outcome that the daughter performs a new role that is absent in the parent (Figure 1.5a). Given that most genes are relatively well adapted to their role, this is likely to be an extremely rare event. However in populations that are sufficiently large, the probability of observing such a beneficial mutation is increased, and depending upon the level of fitness benefit that the mutation provides, there may be very strong selection for the new gene to become fixed (Lynch *et al*, 2001). It is also theoretically possible for neofunctionalisation to occur through drift alone, if, as a result of environmental change, the daughter copy gains some form of fitness advantage (Kimura, 1983).

It should be noted that neofunctionalisation may occur both through changes in coding sequence and through changes in regulatory sequence (Makova & Li, 2003; Huminiecki & Wolfe, 2004; Farré & Albà, 2010; Figure 1.4). Should the daughter copy be duplicated in the absence of its promoter or other regulatory sequence, it may gain a new promoter in its new location, immediately or over time, which may result in rapid neofunctionalisation through change in expression profile. In addition to novel functions, neofunctionalisation can also occur through changes in timing and location of expression. A further path towards neofunctionalisation is the duplication of a neofunctionalised allele – *i.e.* an allele that performs a function that other alleles of the same gene do not,

and subsequent specialisation for that function, leaving the original function(s) to the parent copy, as in the case of insecticide resistant alleles of the acetylcholinesterase gene in *Culex* mosquitoes (Labbe *et al*, 2007). In this model evolution following duplication is expected to be asymmetric, with rapid evolution of the daughter copy, initially through relaxed selection and latterly through positive selection once the advantageous mutation has taken place, while purifying selection maintains the parent copy (Innan & Kondrashov, 2010).

Subfunctionalisation

This term was also introduced by Force *et al* (1999), who developed a formal model based upon observations in the prior decade, which they called the duplication-degeneration-complementation (DDC) model. While their initial DDC model focussed on duplications of chromosomes or entire genomes, it has since become accepted that it can also pertain to duplication of lone genes. Subfunctionalisation generally requires that the parental gene has at least two distinct functions, but may also apply if a single function can be divided in some other way, such as in time or location. Following duplication, if a mutation occurs that results in inactivation of a particular function in one copy, and subsequently a mutation occurs in the other copy that results in the complementary inactivation of another distinct function, then the point is reached where it is necessary to maintain both copies in the genome in order to be able to perform the complete set of functions of the progenitor gene. Such mutations may affect coding sequence directly by disrupting motifs or domains, or may affect regulatory regions thus influencing location or timing of

expression. This model predicts symmetric neutral evolution following duplication until the two copies have become subfunctionalised, at which point they will be maintained by purifying selection (Figure 1.5b). Lynch and Force (2000) have shown that as effective population size increases, subfunctionalisation will become less likely while nonfunctionalisation of the copy that first incurred an inactivating mutation becomes more likely.

An alternative model of subfunctionalisation proposed by Hughes (1994) is that of escape from adaptive conflict (Figure 1.5c). This model assumes that the two functions in the progenitor gene are suboptimal as a result of antagonistic pleiotropy. However, following duplication this pleiotropic constraint will be relaxed and thus each function can become independently optimised in a different duplicate. In this case we would expect to see relatively rapid evolution through the action of positive selection on both copies. An example in nature involving the anthocyanin biosynthetic pathway in the common morning glory has been described (Des Marais & Rausher, 2008).

In reality it is likely that in many instances evolution following duplication will follow a mixture of the paths outlined above (Walsh, 2003; Huminiecki and Wolfe, 2004), and thus some authors have suggested that subfunctionalisation followed by neofunctionalisation may be a prominent mode (He & Zhang, 2005; Rastogi & Liberles, 2005). Using population genetic models, Walsh (2003) has shown that subfunctionalisation is likely to be more important in the case of small effective populations, with neofunctionalisation becoming gradually more prominent as population size increases. Furthermore, in nature many gene duplicates will not be born equal, as assumed by the majority of

these models, due to immediate relocation to a novel chromosomal environment (Cusack & Wolfe, 2007), which is particularly important in the case of duplicates that result from retrotransposition. In such cases, if a duplicate gene survives its birth, then we might expect it to rapidly either neofunctionalise or pseudogenise.

1.5.3 Retroduplication

Retroduplication is the process whereby a mature mRNA species is reverse-transcribed and the resultant cDNA equivalent is reinserted into the genome, more or less at random. As it occurs during meiosis and germ-cell formation, it is predominantly observed in genes that are highly-expressed during this process *i.e.* housekeeping genes and genes that are highly expressed in the germline. In mammals the process is generally mediated by L1 long interspersed nuclear elements (LINEs), which are a family of active retrotransposons that encode a reverse-transcriptase that recognises polyadenylated mRNA species. The resulting daughter retrogenes can be identified by their lack of introns with respect to the parent copy, the presence of a polyA tail, and bordering direct repeat sequences (Figure 1.6). Since retroduplication does not involve the duplication of the proximal promoter region or introns, for a long time retrogenes were assumed to be dead-on-arrival, and thus categorised as processed pseudogenes. However, the first functional retrogene was described by McCarrey and Thomas in 1987. They showed that PGK-2 on chromosome 6 was a functional retrocopy of the X-linked phosphoglycerate kinase gene. They found that the new copy was testis-

specific and actively expressed during spermatogenesis, and thus proposed that it functions as a compensatory response to X-chromosome inactivation during meiosis.

In the post-genomic era many other apparently functional retrogenes with similar characteristics have been described in *Drosophila*, mouse, and human (Betrán *et al*, 2002; Emerson *et al*, 2004; Vinckenbosch *et al* 2006). Vinckenbosch *et al* (2006) report that there are at least 120 functional retrogenes in the human genome, and that there may be as many as one thousand, of which approximately one-quarter have their progenitor copies on the X-chromosome. They also found that intact retrogenes are more likely to be found proximal to, or entirely within, other genes, suggesting that they may often hijack or share corresponding regulatory regions. An alternative manner by which retrogenes may obtain regulatory regions is in cases where alternative transcription start sites are used in the parent copy. In such cases retrocopied UTR regions may provide some degree of promoter activity. As the majority of retrogenes are relocated to a distinct chromosomal environment relative to their progenitors, it is not surprising that, if they survive the move, they tend to show very different patterns of expression, and thus may evolve rapidly at a functional level, and several such cases have been described in detail (see Kaessmann *et al*, 2009 for a review).

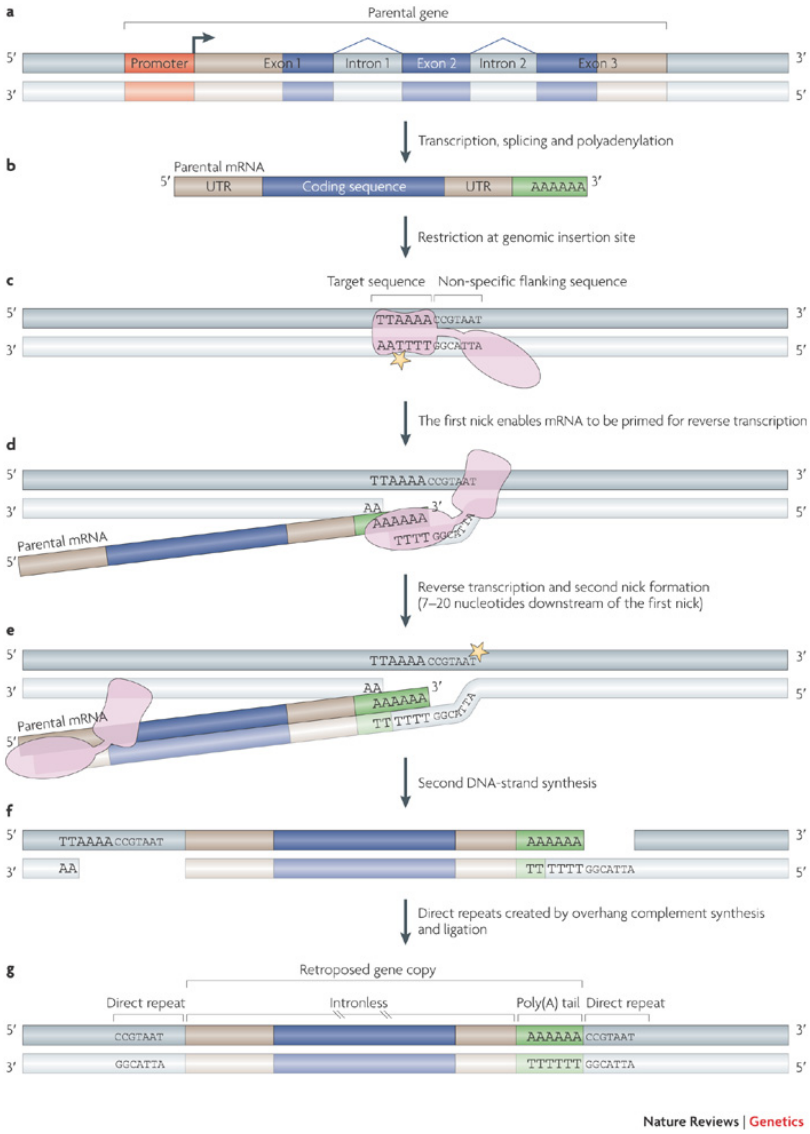


Figure 1.6. Retroduplication. a. Retroduplication is initiated with the transcription of a parental gene by RNA polymerase II. **b.** Further processing of the resulting RNA (splicing and polyadenylation) produces a mature mRNA. **c**. Retroduplication is mediated by the L1 endonuclease domain which creates a first nick (star) at the genomic site of insertion at the TTAAAA target sequence. **d.** This nick enables the mRNA to be primed for reverse transcription by the L1 reverse transcriptase domain **e.** Second-strand nick generation. **f.** Second DNA-strand synthesis. **g.** cDNA synthesis in the overhang regions created by the two nicks. This process creates a duplication of the sequence flanking the target sequence, which is one of the molecular signatures of retroduplication; other signatures include the lack of introns and the presence of a poly(A) tail. The direct repeats and the poly(A) tail degenerate over time, and are therefore usually only visible in recent retrogenes. Abbreviated from Kaessmann *et al* (2009).

1.5.4 Observations to date

A number of different approaches have been taken in the investigation of gene duplication to date. These include analysis of the outcome of whole genome duplication in plants, yeast, and tetrapods (Lynch & Conery, 2000; Davis & Petrov, 2004; Scannell & Wolfe, 2008; Studer *et al*, 2008), analysis of gene families (Huminiecki & Wolfe, 2004; Demuth *et al*, 2006; Hahn *et al*, 2007; Dong *et al*, 2009; Han *et al*, 2009; Chen *et al*, 2010; Farré & Albà, 2010; Ezawa *et al*, 2011), and analysis of identifiable pairs of duplicated genes (Zhang *et al*, 2003; Cusack & Wolfe, 2007). In each case it is likely that the outcomes will be distinct, and thus I will consider them separately.

Whole genome duplication

Scannell & Wolfe (2008) found that proteins that had survived in duplicate following WGD in yeast of the genus *Saccharomyces* tend to evolve faster on average than single copy orthologs in related yeast species that had not undergone WGD. This was in contrast to previous work that had suggested that retained duplicates tended to be more conserved prior to duplication (Davis & Petrov, 2004). Scannell & Wolfe (2008) also found that the two copies tended to evolve at asymmetric rates, though the slow copy was still evolving faster than non-duplicated orthologs in most cases. They describe a burst of evolution following WGD, followed by a gradual reduction in evolutionary rate, though all the duplicate branches they examined remain faster evolving to the present day than equivalent branches in species that have not undergone WGD. Studer *et al* (2008) found no evidence of asymmetry in duplicates, though they were

looking at very old duplicates generated during the vertebrate WGD events, and thus any temporary increase in rate may have completely dissipated by now.

Duplications in gene families

Gene families arise when there have been multiple duplication events of a particular gene and its daughters, leading to a range of genes with typically similar but distinct functions within a genome. Matthew Hahn's group at Indiana University has been one of the most prominent in investigating the evolution of gene families in mammals (e.g. Demuth *et al*, 2006; Hahn *et al*, 2007; Han *et al*, 2009). It should be noted that in their large-scale comparative analyses they also consider gene families of size one *i.e.* where there has been no duplication. In an analysis of 9,900 gene families across human, chimpanzee, mouse, rat and dog, they suggest that at least 10% of families have changed in size in each lineage since their most recent common ancestor (Demuth *et al*, 2006). Using maximum-likelihood estimates, they observed gene gain on the human, mouse, rat, and ancestral rodent branches, and contractions on the remaining branches. They found 164 families, encompassing a wide-range of biological functions including immune defence, transcription, intercellular communication, and metabolism, to be evolving non-randomly. It should be borne in mind that they were using relatively early genome builds, and thus it is not unlikely that many genes were not yet annotated, particularly in chimpanzee and dog, or that they were not classified correctly by the clustering algorithm applied. In addition approximately 60% of this dataset consisted of single-gene families, giving them a mean family size of just under two genes per family. Nevertheless this study provides reasonable evidence of continual birth-and-loss of genes.

Focussing specifically on the subset of primate branches in a subsequent study, Hahn *et al* (2007) suggest that there has been an acceleration of turnover in human and chimpanzee relative to macaque, with gene gain in human and loss in chimpanzee. In a further study focussing on lineage specific duplicates in human, macaque, mouse and rat (mean gene family size ~3), Han *et al* (2009) reported that from a total of ~2400 families they found approximately 10% showed lineage specific evidence of positive selection (*i.e.* dN/dS>1), and paralogs were roughly four times as likely to be evolving under positive selection as single-copy orthologs across these species. In 66 cases where they had evidence of positive selection and could unambiguously identify parent and daughter copies they found that in 80% of the cases it was the daughter copy that was under selection, thus supporting the neofunctionalisation model. Of note, 40% of these 66 cases probably arose through retroduplication.

Dong *et al* (2009), performed an in-depth study of olfactory receptor genes, which comprise the largest mammalian gene family, in primates. They found that they tend to have 300-400 members per species, which is approximately one-third of the number that has been reported in mouse and rat, a fact that has been ascribed to large-scale loss following the development of trichromatic colour vision in primates (Gilad *et al*, 2004). Dong and colleagues found wide-scale birth and loss, with approximately 30% change in composition of olfactory receptor repertoire between human and chimpanzee since their divergence just 6Mya. While it should be noted that correct cross-species ortholog identification in such a large family is challenging, the central finding of high turnover in gene number in this family is likely to be correct. Further it

has subsequently been shown that the olfactory receptor family includes segregating pseudogenes in human *i.e.* genes which have both functional and non-functional alleles, illustrating that evolution towards pseudogenisation is not necessarily a one-way street (Hinkley & Ismaili, 2012). Chen *et al* (2010) looked at conservation of size of gene families shared between human, chimpanzee and macaque, and observed that multi-gene families whose size had not been conserved across all three species were evolving faster than those families where size has been conserved across them. Families constrained to one copy across the three species, which comprised 80% of the dataset, fell in between the two other classes, and they also observed that size-conserved families had a higher proportion of essential genes, having higher and broader expression levels.

Duplogs

An alternative approach to investigating families of genes is to look at pairs of young duplicates (duplogs) individually. Initial investigations of this form in vertebrate species found very little evidence of divergence in evolutionary rate between within-species paralogs. Robinson-Rechavi and Laudet (2001) reported just 4 out of 19 mammalian pairs were evolving significantly differently from one another at $p<0.05$, and this dropped to zero following correction for multiple testing, while Kondrashov *et al* (2002) reported just 2 out of 49 cases analysed in mammals. However, Zhang *et al* (2003) investigated 250 pairs of young duplicates in the human genome and found that 10-20% showed significant differences in terms of dN/dS between the copies. They also found that the fast copy had substitutions spread across its sequence while the slow copy had more uneven patterns, suggesting relaxation in the former and

constraint in the latter, and in seven cases there was evidence for positive selection (dN/dS significantly greater than 1). In an analysis of 147 recent rodent duplicates, Cusack and Wolfe (2007) found that both relocation and retroduplication were independently associated with asymmetry in evolutionary rates and that approximately 30% of each of these classes of pairs were evolving asymmetrically, whereas pairs that remained in tandem duplication did not display significant asymmetry.

Divergence of expression in duplicates

Another question of interest regarding paralogs is how they may differ in expression. Early studies in yeast using microarray data indicated that paralogs tend to show divergence in expression levels, that this divergence can arise relatively rapidly following duplication, and that it is correlated with sequence divergence (Wagner, 2000; Gu *et al*, 2002). Subsequent analyses in humans (Makova & Li, 2003) found similar results, but more rapid divergence per generation, and that paralog expression patterns tend to become more specialised as a gene family grows in size (Huminiecki & Wolfe, 2004). Huminiecki and Wolfe also found that for pairs of young gene duplicates, in most cases both duplicates had diverged away from the predicted ancestral state, generally in a pattern that would suggest subfunctionalisation, though they also found some evidence of neofunctionalisation in expression pattern as well. However they note that they typically observed a degree of divergence in expression pattern between one-to-one orthologs of human and mouse as well, thus questioning the reliability of predictions of an ancestral state of expression. Cusack & Wolfe (2007) also found evidence of a small degree of expression divergence in distantly separated pairs, and that retrogenes as a group had

significantly narrower expression breadth than their parent copy. Farré & Albà (2010) found that gene duplication in rodent gene families is frequently associated with a reduction in expression breadth and intensity of individual paralogs, but cases fitting the classical model of neofunctionalisation were rare.

1.6 Data Collation

The data used in the analyses undertaken here derives from collaborative international genome sequencing projects, in particular those of human (Lander *et al*, 2001), mouse (Waterston *et al*, 2002), rat (Gibbs *et al*, 2004), macaque (Gibbs *et al*, 2007), opossum (Mikkelsen *et al*, 2007), and cow (Elsik *et al*, 2009). These genomes were all published, and hence publicly available, before I started this project. However, initial publication of a complete vertebrate genome has come to refer to completion of a *draft* genome, generally consisting of ~6-7-fold coverage[5], and not to completion of a *finished*[6] genome. Of the species considered here, so far only the human and mouse genomes have been finished (International Human Genome Sequencing Consortium, 2004; Church *et al* 2009), and with the exception of the important model vertebrates zebrafish and rat, it is unlikely that other vertebrate genomes will reach such a stage of completion in the near future. Cheaper sequencing technologies may help address this shortfall at some point (English *et al,* 2012), but in the meantime the quality of the majority of vertebrate genomes remains in draft form at best,

5 Coverage refers to the average number of times each specific nucleotide has appeared in a read during the sequencing process. Essentially it approximates to the amount of sequence generated by the project divided by the length of the genome being sequenced.

6 There is no hard definition of a finished genome, but for human and mouse it was taken to mean 99% euchromatin coverage with 99.99% base calling accuracy.

and in most cases is currently limited to low-coverage (2-fold) genomes (Lindblad-Toh *et al*, 2011) which are unsuitable for the types of analyses undertaken here due to the inflated sequencing error rate (Hubisz *et al*, 2011).

Production of a draft genome can be separated into at least four distinct stages: whole genome sequencing to the depth of coverage required; assembly of sequencing fragments into contigs[7]; mapping of these contigs to some form of reference genome or physical map; gene identification and prediction. The depth of coverage will directly affect the assembly and mapping stages, which will in turn affect the accuracy of gene prediction. Thus the higher the coverage, the better the quality of the resulting genome. In order to reach the accuracy required of a finished genome, somewhere in excess of 30-fold coverage is necessary.

Following completion of sequencing, the next stage is genome assembly. Genomes are not sequenced in a linear fashion, but broken into overlapping pieces which are in turn broken into shorter pieces which are then sequenced individually. In the case of the human and mouse genomes the segments were cloned into bacterial artificial chromosome (BAC) libraries, each of which held approximately 150kb of DNA. BACs were sequenced individually using capillary sequencing, and mapped through hybridisation to a particular chromosome region, so they could then be aligned and merged together to build complete chromosomes, which range from approximately 50Mb to 250Mb in length (Lander *et al*, 2001; Waterston *et al*, 2002).

7 A contig is a contiguous sequence of bases that has been constructed by aligning overlapping reads and merging them together to provide a consensus sequence.

However, the establishment and maintenance of a BAC library is relatively expensive, and thus most eukaryote genomes have since been sequenced using the whole genome shotgun (WGS) approach pioneered by Craig Venter and colleagues at Celera for the sequencing of the *Drosophila* genome (Adams *et al*, 2000), and during their competition with the publicly-funded human genome project (Venter *et al*, 2001). This involves breaking the genome into tiny chunks for sequencing, and latterly trying to reconstruct the genome from these individual reads. However, complete genome reconstruction following WGS is very difficult, and currently unfeasible for eukaryotic genomes, without some from of physical map on which to hang the contigs generated.

Unfortunately the WGS process does not lead to a uniform distribution of reads and thus some loci will be sequenced many times by chance, while others will be sequenced very few times, if at all. In addition, certain regions of genomes often prove more difficult to sequence, and quality of sequence also varies between reads, resulting in patches of the genome where the sequence quality is low. While sequencing software packages provide a quality score for each sequenced base, as primary output they report the best guess at each position and thus it may not be immediately apparent to the end-user that a particular section of sequence may be unreliable. The difficulty of genome assembly is compounded further in genomes that have many regions of repetitive sequence, and when using second generation sequencing technologies where read length is currently substantially shorter than that produced by the capillary sequencing techniques used to build earlier genomes (i.e. approximately 50-400bp versus 600-1000bp per read).

Following assembly and mapping, the final step in producing a biologically useful complete genome sequence is the identification of the functional units[8]. A number of algorithms have been developed that can sift through a genome sequence automatically and identify regions that appear to have the characteristics of being part of a protein-coding gene i.e. plausible start codons, splice site motifs, stop codons, and 5' and 3' untranslated regions where appropriate (Burge & Karlin, 1997; Birney *et al*, 2004; Curwen *et al*, 2004; Gross *et al* 2007). These algorithms often incorporate sources of experimental evidence, particularly in the form of expressed sequence tags[9], when building gene predictions. However, while these algorithms work relatively well on the whole, when one begins to look in more detail at individual predicted gene structures, one often finds apparent inconsistencies, sometimes due to limitations in the algorithm and sometimes due to underlying sequencing or assembly errors. As a biologist we must then ask ourselves if we believe these inconsistencies to be real, in which case they may be of biological interest, or if they are merely artefacts of the gene prediction algorithm, in which case they should be ignored.

An excellent illustration of these issues was provided by Florea *et al* (2011) who built two successive genome assemblies for cow using the same initial raw reads. Their second version involved more accurate filtering to remove vector sequence and reads identified as originating from bacterial contaminants. This

8 Formerly functional units were restricted to genes, but now include many species of non-coding RNAs, and conserved non-coding elements.

9 Expressed sequence tags are short sequences of complementary DNA that have been generated from the sequencing of mRNA species and thus represent portions of sequence from expressed genes.

significantly improved the accuracy of the assembly and reduced the fraction of unmapped sequences from ~8% to less than 0.3%. Of particular note, they found that only 62% of predicted transcripts had completely preserved exon-intron structures between their two assemblies.

The most important publicly accessible repositories of genome scale data are Ensembl from the European Bioinformatics Institute at Hinxton in the UK and the UCSC Genome browser, from the University of California at Santa Cruz. These repositories provide overlapping and complementary data, and the choice of which to use will depend upon the biological questions being addressed.

The UCSC genome browser (http://genome.ucsc.edu, Meyer *et al* 2013) currently contains data pertaining to 39 vertebrates and 24 invertebrates. It has a sequence-based perspective providing whole genome pairwise alignments between human and each of the other species, and incorporating sequence tracks from many diverse sources of experimental results for easy online visualisation. Ensembl (http://www.ensembl.org, Flicek *et al*, 2013) focusses primarily on the curation of coding sequence of chordate genomes, currently numbering 61 (Ensemble release 71, April, 2013), together with data from the finished genomes of three important model organisms in *C.elegans*, *D.melanogaster* and *S.cerevisiae* which are valuable for use as outgroups in comparative analyses. It produces automated gene sets for all species, together with variation data, annotation of regulatory regions, and multiple alignments for various subsets thereof. New releases of the database are produced every three to four months, and species gene sets are updated and complete new genome builds produced, as and when new data become available.

Ensembl also undertakes homology prediction through application of its *Compara* pipeline, endeavouring to correctly identify orthologous genes across species and paralogous genes within species utilising sequence similarity, clustering, and phylogenetic tree reconstruction to achieve this goal (Vilella *et al*, 2009). Ensembl thus classifies orthology between homologous genes of any two particular species into one of three categories:

1) one-to-one orthologs: true, simple relationship – the same gene in each species[10].
2) one-to-many orthologs: where the original ortholog in one of the two species has undergone duplication at some point in its evolutionary history, and thus there are two or more paralogs in this species that map to just one unduplicated ortholog in the other.
3) Many-to-many orthologs: here duplication of the original orthologous genes has taken place in both species independently and thus we have families of paralogous genes that are related to each other within, and between, the species concerned.

10 It should be noted that genes that are part of a large family that formed as a result of historic duplications may still be classified as one-to-one by Ensembl when there is a sufficient degree of certainty regarding the relationship.

2 RESULTS

This section consists of two published articles

The first paper, *Sequence shortening in the rodent ancestor,* consists of an in-depth analysis of indels in four mammalian species and their ancestral branches. Utilising a relatively new alignment algorithm, PRANK, we compare the frequency of indels in neutrally-evolving ancestral repeat sequence with that for protein coding one-to-one orthologs. We find a correlation between indel incorporation and point mutation, and that selection acts more strongly against the incorporation of insertion than deletions in proteins. In contrast to previous reports, we do not observe a universal bias towards deletions. However we do observe a significant deletional bias in the rodent ancestral branch, equating to a loss of approximately 2.5% of syntenic region in the ancestor of mouse and rat.

The second paper, *Accelerated evolution after gene duplication: a time-dependent process affecting just one copy*, investigates evolution following gene duplication in rodents. We show that following duplication, there is a general trend for the original copy to continue evolving at the pre-duplication rate, while the new copy shows marked acceleration over a period of 4-12MY, before gradually returning to pre-duplication rates. We find evidence that positive selection plays a significant part in this process in many cases, and that gene duplication is often accompanied by divergence in tissue expression patterns, providing support for the neofunctionalisation model.

2.1 Sequence shortening in the rodent ancestor

Title: Sequence shortening in the rodent ancestor

Authors: Steven Laurie, Macarena Toll-Riera, Núria Radó-Trilla, and M. Mar Albà

Published in: Genome Research (2012), 22: 478-485

Full text: http://genome.cshlp.org/content/22/3/478

doi: 10.1101/gr.121897.111

Abstract

Insertions and deletions (indels), together with nucleotide substitutions, are major drivers of sequence evolution. An excess of deletions over insertions in genomic sequences—the so-called deletional bias—has been reported in a wide range of species, including mammals. However, this bias has not been found in the coding sequences of some mammalian species, such as human and mouse. To determine the strength of the deletional bias in mammals, and the influence of mutation and selection, we have quantified indels in both neutrally evolving noncoding sequences and protein-coding sequences, in six mammalian branches: human, macaque, ancestral primate, mouse, rat, and ancestral rodent. The results obtained with an improved algorithm for the placement of insertions in multiple alignments, Prank+F, indicate that contrary to previous results, the only mammalian branch with a strong deletional bias is the rodent ancestral branch. We estimate that such a bias has resulted in an ~2.5% sequence loss of mammalian syntenic region in the ancestor of the mouse and rat. Further, a comparison of coding and noncoding sequences shows that negative selection is acting more strongly against mutations generating amino acid insertions than against mutations resulting in amino acid deletions. The strength of selection against indels is found to be higher in the rodent branches than in the primate branches, consistent with the larger effective population sizes of the rodents.

2.2 Accelerated evolution after duplication: A time-dependent process affecting just one copy

Title: Accelerated evolution after duplication: A time-dependent process affecting just one copy

Authors: Cinta Pegueroles*, Steve Laurie*, and M. Mar Albà

Published in: Molecular Biology and Evolution (2013), Advanced Access online May 8th

Full text: http://mbe.oxfordjournals.org/content/early/2013/05/17/molbev.mst083.long

doi: 10.1093/molbev/mst083

Abstract

Gene duplication is widely regarded as a major mechanism modeling genome evolution and function. However, the mechanisms that drive the evolution of the two, initially redundant, gene copies are still ill defined. Many gene duplicates experience evolutionary rate acceleration, but the relative contribution of positive selection and random drift to the retention and subsequent evolution of gene duplicates, and for how long the molecular clock may be distorted by these processes, remains unclear. Focusing on rodent genes that duplicated before and after the mouse and rat split, we find significantly increased sequence divergence after duplication in only one of the copies, which in nearly all cases corresponds to the novel daughter copy, independent of the mechanism of duplication. We observe that the evolutionary rate of the accelerated copy, measured as the ratio of nonsynonymous to synonymous substitutions, is on average 5-fold higher in the period spanning 4–12 My after the duplication than it was before the duplication. This increase can be explained, at least in part, by the action of positive selection according to the results of the maximum likelihood-based branch-site test. Subsequently, the rate decelerates until purifying selection completely returns to preduplication levels. Reversion to the original rates has already been accomplished 40.5 My after the duplication event, corresponding to a genetic distance of about 0.28 synonymous substitutions per site. Differences in tissue gene expression patterns parallel those of substitution rates, reinforcing the role of neofunctionalization in explaining the evolution of young gene duplicates.

Pegueroles C, Laurie S, Alba MM. **Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. Supplementary material.** Mol Biol Evol. 2013 Aug;30(8):1830-1842.

3 DISCUSSION

3.1 Methodology Applied

Bioinformatics is a wide-ranging and interdisciplinary field, encompassing many different techniques. However, with the exception perhaps of modelling, it tends to be observational and analytical in nature, rather than experimental, thus distinguishing it from the majority of work undertaken in a traditional *wet* lab. Nevertheless both endeavours are intrinsically linked; wet lab biologists conduct the experiments that generate the data that bioinformaticians analyse, and the results of these analyses help to generate new hypotheses requiring testing by further experiments in the laboratory, and so the circle continues.

This thesis represents a purely bioinformatic body of work, consisting of the analysis of data that were generated by countless other people, but unfortunately lacking in the opportunity for dialogue, discussion and feedback with these individuals. Hence I have naturally had to make a number of assumptions during the course of this work, some of which are clearly sensible and robust, but others of which may be open to criticism. These assumptions include that the sequence being analysed is biologically correct, that the genes being compared are truly equivalent, and that the programs that have been used to analyse and process the raw data have performed reliably. As the work undertaken here was primarily gene-centred in nature, I used Ensembl (Flicek *et al*, 2013) as the primary source of protein sequence data for my analyses. However information from the UCSC Genome Browser (Meyer *et al*, 2013) was also used where appropriate, particularly for the analyses utilising non-coding ancestral repeat sequences.

3.1.1 Raw Data Quality Control

When undertaking comparative genomics analyses we want to be confident that the sequences we are comparing are truly homologous. Therefore, for the indel analysis project that formed the first part of this thesis we chose to work only with genes that are classified as one-to-one orthologs by Ensembl across all five species in the dataset. This immediately reduced the number of possible sets of genes for analysis by approximately 50%, including all cases where there was other than one-to-one orthology, and cases where no ortholog was found in a particular species. While there are certain to be cases where there is no human ortholog observed in other mammalian species, there are also likely to be cases where homology relationships have been misclassified and where particular genes may not yet have been correctly identified in the genomes concerned.

Ensembl typically lists more than one coding transcript per gene for human (mean of 9) and mouse (mean of 4), but the number of transcripts identified in the other species used here is currently severely limited (mean of between 1.2 and 2, Ensembl Release 71, April 2013). Since homology in Ensembl is determined at the level of the gene and not at that of the transcript, and because many transcripts will not yet have been described in species other than human and mouse, we chose to take the longest available coding transcript as representative of the gene under consideration. This means that we will not always have been aligning fully homologous sequences, which we addressed by applying a number of filters based upon similarity in overall length and similarity of sequence at the exon level. As a result of this process, we became

aware of a number of cases where parts of the amino acid sequence were unidentified in particular proteins, indicating poor quality of underlying sequence data, particularly affecting proteins from macaque, and to a lesser extent those of rat.

This observation led us to investigate raw sequence quality scores in more detail. Sequence analysis software packages provide a measure of the quality of each individual base in the form of a Phred of Q-score, or equivalent, which is a negative-log-score index where scores of 40 or more represent the *de facto* gold standard of a base-calling error rate lower than $1x10^{-4}$. Unfortunately such scores were only available for macaque and cow, but nevertheless they confirmed what we suspected from visual inspection of the alignments, which was that the macaque sequences were relatively poor. Overall, eleven percent of all macaque exons in this study had at least one nucleotide with a score less than 40, compared to just five percent in cow. However, it would appear that there have also been problems with the macaque assembly and gene prediction, as poorly aligned macaque exons accounted for more than their fair share of post-alignment filtering based upon exon sequence conservation (see Section 3.1.3).

3.1.2 Choice of Alignment Algorithm

Once we had established our initial clean dataset, the next step was to decide upon which multiple alignment program to use. This is a key decision since different algorithms produce different alignments, thus affecting downstream deductions (Golubchik *et al*, 2007; Wong *et al*, 2008). Currently there are four MSA algorithms that are relatively commonly cited in the literature, ClustalW

(Thompson *et al*, 1994) , T-Coffee (Notredame *et al*, 2000), MAFFT (Katoh *et al*, 2002), and MUSCLE (Edgar, 2004), and one new algorithm that is gaining popularity, PRANK (Löytynoja & Goldman, 2008). ClustalW is by far the oldest and still most commonly used, while PRANK, being the newest is the least used thus far (see Table 1.1). Comparisons of the remaining three algorithms have shown that they produce similar results with regards to accuracy, but MAFFT is by far the fastest (Golubchik *et al*, 2007; Thompson *et al* 2011). Since PRANK was developed specifically with the goal of reducing over-alignment of insertions with non-homologous sequences later in the progressive alignment process (Löytynoja & Goldman, 2005), it was a natural choice for this study. Thus we chose to compare alignments generated by ClustalW (the traditional standard), MAFFT (the fastest of the rest), and PRANK. Alignments are traditionally compared based upon the number of columns that are identical, or the numbers of pairs of residues that are shared (Thompson *et al*, 1999a). However, neither of these metrics can accurately measure over-alignment, which results in incorrect gap-merging, so they do not represent perfect measures of alignment quality.

As we were specifically interested in the gaps in the alignments, and knew from preliminary analyses that there were annotation errors in the dataset, we chose to examine the number of complete exons that did not align with any other exon in the alignments produced by the different alignment programs in order to compare how each of these algorithms performed in identifying putative insertions. By mapping the position of all individual exons within each alignment, we found that PRANK far outperformed ClustalW, and significantly outperformed MAFFT in this respect, being able to identify and separate out

non-homologous exons, thus markedly reducing over-alignment of non-homologous exons (see Figure 1.3 and Supplementary Table 1 in Section 2.1.1). Identification of a non-homologous exon in this manner is equivalent to identification of an insertion in a particular sequence, and thus these results clearly indicate that PRANK is best in this respect. Indeed, while the three algorithms identify largely equivalent numbers of deletions in the dataset, PRANK identified significantly more insertion events throughout (see Supplementary Table 3 in Section 2.1.1). Thus, while PRANK, still does not produce perfect alignments, it produces more reliable alignments with datasets of this type, and our findings concur with those of others who have shown that PRANK is currently the best MSA algorithm for use in comparative analyses where tests for positive selection will be applied to the alignments generated (Mallick *et al*, 2009; Schneider *et al*, 2009; Fletcher & Yang, 2010; Markova-Raina & Petrov, 2011; Jordan & Goldman, 2012).

3.1.3 Post-alignment Filtering

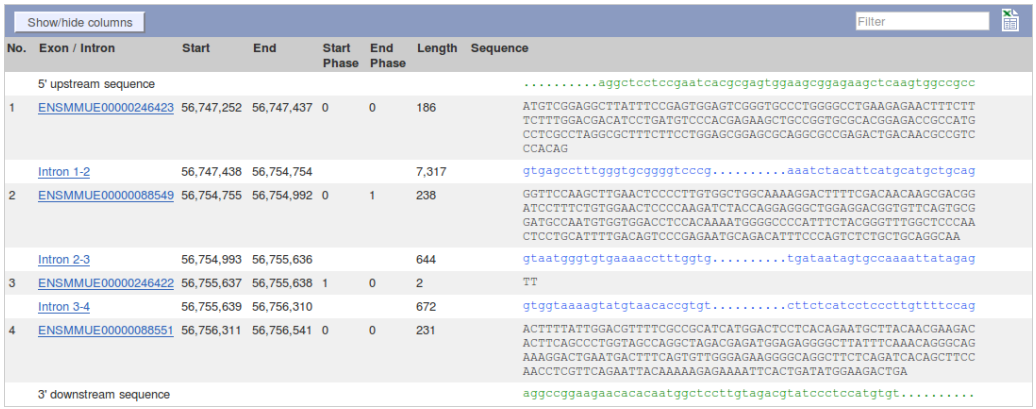
Ideally one would not have to perform any post-alignment filtering. However, often it is only once alignments are generated that one realises that there is something wrong with the sequences being aligned. Attempting to measure rates of sequence evolution is always an underlying goal in these analyses, which is known to be highly sensitive to alignment errors (Wong *et al*, 2008; Mallick *et al*, 2009; Schneider *et al*, 2009; Fletcher & Yang, 2010; Markova-Raina & Petrov, 2011; Yang & dos Reis, 2011). Therefore, in order to have the best subset of sequences for measuring evolutionary rate, we chose to remove cases where there were badly defined exons. We set a threshold of exon

similarity of at least 50% identity to achieve this, which is generally conservative since it is estimated that 70% of amino acids are conserved between human and mouse (Waterston *et al*, 2002).

Furthermore, only after we generated our alignments did we become aware of the issue of orphan and truncated/extended exons, which were particularly prominent in macaque once again. Following visual inspection it was clear in many cases that the gene prediction algorithm had misidentified splice sites. We addressed this issue by mapping the position of all exons onto the alignments in order to directly identify such events less they be erroneously counted as indels. We discounted all observations of indels which were immediately adjacent to an exon boundary as they were found to typically represent annotation errors, and such sites were found to often result in false-positive signals of selection in an analysis of *Drosophila* species (Markova-Raina & Petrov, 2011). It would appear that in cases where the gene prediction algorithm applied to the macaque did not find evidence for a truly homologous exon, it instead identified similar sequences from within intervening introns and labelled them as exons (see Figure 3.1 for an example). As far as we are aware, this level of attention to detail has not been applied in any study of this nature previously.

Show/hide columns | Filter

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence						.....ggcgcccccacctaccgctccccagccgtgtcccgctgctctaaatctgc
1	ENSE00001401750	58,426,298	58,426,691	-	0	394	AGACTTGATCGATTGCTTCTGCCTGGGCGGTACCGCCCGAATTGACTGCTCCTGTCTGAT GCGTCCCCGGGCGCGGGAAACGAGTTTCAATCCACTTTCCTGACCCCAACCATCCTGCCC AGTCTCCGCTTCCCCGTCTTGTACACCCCTAACTCCTGAGGCTCCTCCGAATCACGCGAG TGGAAGCGGAGAAGCTCAAGTGGCCGCCATGTCAGAGGCTTATTTCCGAGTGGAGTCGGG TGCGCTGGGGCCTGAGGAGAACTTTCTTTCTTTGGACGACATCCTGATGTCCCACGAGAA GCTGCCGGTGCGCACGGAGACCGCCATGCCTCGCCTTGGCGCTTTCTTCCTGGAGCGGAG CGCAGGCGCCGAGACTGACAACGCGGTCCCACAG
	Intron 1-2	58,426,692	58,437,001			10,310	gtgagcctttgggtgcggggtcctg.....aaatctacattcatgcatgctgcag
2	ENSE00001218868	58,437,002	58,437,235	0	0	234	GGTTCCAAGCTTGAACTACCCTTGTGGCTGGCAAAAGGACTTTTTGACAACAAGCGACGG ATCCTTTCTGTGGAACTCCCCAAGATCTACCAAGAGGGTTGGAGGACTGTGTTCAGTGCA GATCCCAATGTGGTGGACCTCCACAAAATGGGGCCCCATTTCTACGGGTTTGGCTCCCAG CTCCTGCATTTTGACAGTCCCGAGAATGCAGACATTTCCCAGTCTCTGCTGCAG
	Intron 2-3	58,437,236	58,438,402			1,167	gcaagtaatgggtgtgaaaacctgt.....cttgtcttactcccttgtttcccag
3	ENSE00001829311	58,438,403	58,440,048	0	-	1,646	ACTTTTATCGGACGTTTTCGCCGCATCATGGACTCCTCACAGAATGCTTACAACGAAGAC ACTTCAGCCCTGGTAGCCAGGCTAGACGAGATGGAGAGGGGCTTATTTCAAACAGGGCAG AAAGGACTGAATGACTTTCAGTGTTGGGAGAAGGGGCAGGCTTCTCAGATCACAGCTTCC AACCTCGTTCAGAATTACAAGAAGAGAAAATTCACTGATATGGAAGACTGAAAGCCGGAA GAACACAGAATGGCTCCTCACAGACGTATCCCTCCGTGTGTCCTTGATAGGAGCTGGTTG ACCTTGTACAGAACCAGAATCCTGTCCCATTTCATGGCTTATTTCCTGTGGCCATAGAGA ATTATAGGGAACTGGACATGCTGGAGGATGTGGGTGTCCCTGGCTCTGTGAGTCTTCCAG GACCGTCCCACCCTGCTGACCCACAGCCCAGGCCCTTTAACCCAAGAACCCATGGCCAAG GAGAAATCAAAGTCCTTCCTAAATAAGAATCACTGCCATATAATATATCACAGTAGAGTT



Show/hide columns | Filter

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence						.....aggctcctccgaatcacgcgagtggaagcggagaagctcaagtggccgcc
1	ENSMMUE00000246423	56,747,252	56,747,437	0	0	186	ATGTCGGAGGCTTATTTCCGAGTGGAGTCGGGTGCCCTGGGGCCTGAAGAGAACTTTCTT TCTTTGGACGACATCCTGATGTCCCACGAGAAGCTGCCGGTGCGCACGGAGACCGCCATG CCTCGCCTAGGCGCTTTCTTCCTGGAGCGGAGCGCAGGCGCCGAGACTGACAACGCCGTC CCACAG
	Intron 1-2	56,747,438	56,754,754			7,317	gtgagcctttgggtgcggggtcccg.....aaatctacattcatgcatgctgcag
2	ENSMMUE00000088549	56,754,755	56,754,992	0	1	238	GGTTCCAAGCTTGAACTCCCCTTGTGGCTGGCAAAAGGACTTTTCGACAACAAGCGACGG ATCCTTTCTGTGGAACTCCCCAAGATCTACCAGGAGGGCTGGAGGACGGTGTTCAGTGCG GATGCCAATGTGGTGGACCTCCACAAAATGGGGCCCCATTTCTACGGGTTTGGCTCCCAA CTCCTGCATTTTGACAGTCCCGAGAATGCAGACATTTCCCAGTCTCTGCTGCAGGCAA
	Intron 2-3	56,754,993	56,755,636			644	gtaatgggtgtgaaaacctttggtg.....tgataatagtgccaaaattatagag
3	ENSMMUE00000246422	56,755,637	56,755,638	1	0	2	TT
	Intron 3-4	56,755,639	56,756,310			672	gtggtaaaagtatgtaacaccgtgt.....cttctcatcctcccttgttttccag
4	ENSMMUE00000088551	56,756,311	56,756,541	0	0	231	ACTTTTATTGGACGTTTTCGCCGCATCATGGACTCCTCACAGAATGCTTACAACGAAGAC ACTTCAGCCCTGGTAGCCAGGCTAGACGAGATGGAGAGGGGCTTATTTCAAACAGGGCAG AAAGGACTGAATGACTTTCAGTGTTGGGAGAAGGGGCAGGCTTCTCAGATCACAGCTTCC AACCTCGTTCAGAATTACAAAAAGAGAAAATTCACTGATATGGAAGACTGA
	3' downstream sequence						aggccggaagaacacacaatggctccttgtagacgtatccctccatgtgt.....

Figure 3.1. Ensembl exon and intron prediction for the GINS3 gene in human (above) and its macaque ortholog (below). These screenshots from Ensembl illustrate that while the human protein has three exons, the macaque protein appears to have four exons, one of which is a biologically highly unlikely two nucleotides long. This is almost certainly an error of the gene prediction algorithm used for the macaque, most likely as a result of the rare AG-GC splice junction between exon 2 and intron 3 (correctly identified in human). This junction is missed in the macaque prediction which also extends exon 2 by four nucleotides, resulting in the incorrect addition of two amino acids to the macaque protein sequence, which would be interpreted falsely as an insertion had such cases not been identified by our pipeline. Note also the annotation of 5' and 3' untranslated regions in human (in purple), which are missing from the macaque annotation. See also Figure 1.2 for the corresponding MSA for this gene.

3.2 Indels in the evolution of mammalian proteins

Many different methodologies have been applied in studies that have attempted to investigate indels thus far. The majority of early studies, and some more recently, have been performed using ClustalW (e.g. Ophir & Graur, 1997; Tian *et al*, 2008, Chen *et al*, 2009, McDonald *et al*, 2011) or MUSCLE (Kuo & Ochman, 2009, Wang *et al*, 2009). However, with the arrival of the UCSC multi-track alignments generated using MultiZ (Blanchette *et al*, 2004), many investigators have chosen to take these alignments as their starting point (e.g. Chen *et al*, 2007; Kvikstad *et al*, 2007; Messer & Arndt, 2007; Tanay & Sigia, 2008; Kvikstad *et al*, 2009) or have used BlastZ which underpins MultiZ *(*Schwartz *et al*, 2003; Taylor *et al*, 2004; Fan *et al*, 2007; Chen *et al*, 2009). Similar variety is found in terms of the size of the indel considered to be of interest, generally ranging from 1-100bp, but sometimes of any length. There has also been wide variation in the type of filtering applied regarding sequence conservation or quality surrounding gaps in order for putative indels to be considered *bona fide* (Taylor *et al*, 2004; Kvikstad *et al*, 2007 & 2009; Leushkin *et al*, 2012), or that gaps be separated by some minimum distance from one another (Messer & Arndt, 2007; Tanay & Sigia, 2007; Tian *et al*, 2008; McDonald *et al*, 2011).

This wide range in methodologies illustrates the difficulty inherent in being confident that observed events are real. We chose not to apply any filtering based upon nearby gaps, nor sequence conservation, as each of these requires further assumptions, which are unnecessary at best and may be incorrect at worst (Wong *et al*, 2008; Jordan & Goldman, 2012). We did not investigate

clustering of indel events directly, but given that we observe indels to be more common in low-complexity regions in proteins and that they are particularly enriched in regions of amino acid tandem repeats, we would expect to observe some degree of clustering. For this reason application of a filter based upon proximity to other gaps seems unjustified as does requiring an arbitrary degree of sequence conservation in the region proximal to the indel. Multiple pairwise alignments such as those generated by BlastZ lose information, particularly with respect to gap placement, that true MSA algorithms can use to better identify evolutionary history. Utilising the algorithm implemented in PRANK together with the known phylogenetic tree to guide identification of insertions largely resolves this issue. Thus, though we extracted ancestral repeat alignments from the UCSC Genome Browser which were generated using MultiZ, we then realigned them using PRANK.

Nevertheless, one case for which there is no foolproof way to accurately identify the number of indel events that have occurred is in regions of tandem amino acid repeats. In such regions multiple events will often have occurred in different lineages as indicated by overlapping gaps in the alignment. Some investigators have attempted to enumerate such instances by applying parsimony, or simply merging overlapping gaps into one observation (Cooper *et al*, 2004; Chen *et al*, 2009; Wang *et al*, 2009), neither of which is entirely satisfactory. The alternative approach is to discard regions containing multiple overlapping events from further analysis (Chen *et al*, 2007; Kvikstad *et al*, 2007 & 2009; Tian *et al*, 2008). Here we chose to apply the latter approach, and hence the absolute number of events will be somewhat higher than reported here, but we trust that the relative frequencies are, on the whole, correct.

One of the most interesting results to come out of this research is the observation that insertions appear to be more strongly selected against than deletions in coding sequence in all branches apart from macaque. To our knowledge no such bias has been described elsewhere. It is possible that this may be some form of artefact related to the ancestral repeat sequences used to estimate background genomic frequencies of insertion and deletion, which may be more prone to insertions for some reason. It is also possible that if we were able to include figures for the tandem-repeat regions that we had to discount, this observation may dissipate somewhat, since these regions are known to exhibit a tendency to increase in length. Nevertheless the results presented here suggest that in general these proteins tolerate deletions better than insertions, and further investigation should be undertaken to understand why this should be the case.

As previously described by others (see Section 1.3.2), we also observe a strong correlation between indels and substitution rates. There are two, non-mutually exclusive, possible explanations for this observation. Firstly, certain regions of proteins such as loops and disordered regions may be less constrained by selection and thus more likely to admit both substitutions and indels, or alternatively one may lead to the other. Tian *et al* (2008) found a significant elevation in nucleotide divergence in the first few bases adjacent to an indel and hence suggested that indels are mutagenic, leading to subsequent point mutations in the vicinity of the initial indel event. Leushkin *et al* (2012) have taken this idea further and suggested that the elevated substitution rates surrounding indel events is driven by positive selection acting in a manner to

compensate for the assumed deleterious effect of indel incorporation. Interestingly, they suggest that insertions in coding regions are more disruptive than deletions, in agreement with our findings. They come to this conclusion as they find that insertions in *Drosophila* proteins are accompanied by one amino acid change in the surrounding sequence on average, while deletions are associated with five changes, and provide evidence that this is a result of positive selection. In contrast, McDonald *et al* (2011) have suggested that the underlying sequence in which indels tend to occur is often prone to inducing replication fork stalling resulting in increased nucleotide divergence due to recruitment of repair polymerases with lower levels of fidelity. At the opposite end of the spectrum, Tóth-Petróczy & Tawfik (2013) suggest that accumulation of neutral substitutions precedes indel events. Unfortunately neither Tian *et al* (2008), nor McDonald *et al* (2011), nor Tóth-Petróczy & Tawfik (2013) attempted to separate insertions from deletions in their analyses, so we cannot compare these results with those of our own or those of Kvikstad *et al* (2007), who proposed that different molecular mechanisms lead to the generation of insertions and deletions. Clearly further investigation of this matter is warranted.

Another key result is that we do not observe a universal deletion bias. Indeed we see no deletion bias whatsoever in the non-coding ancestral repeat sequences in mouse and only marginal significance in human ($p=0.03$), while observing the reverse trend in the primate ancestral branch. It is interesting to note that the absence of evidence for bias occurs in the two finished genomes, suggesting that previous observations of widescale deletion bias may, at least in some cases, be the result of artefacts in alignment methodology. These results

suggest that we can confidently refute Kuo and Ochman's (2009) assertion of a universal deletion bias.

It is interesting that we observe differences in the frequency of insertion and deletion across these mammalian branches. Such differences have also been observed in other eukaryotes (Taylor *et al*, 2004; Kvikstad *et al*, 2007; Tóth-Petróczy & Tawfik, 2013) and thus may reflect differences in life-history traits, underlying mutation dynamics, and perhaps molecular machinery. At this point we do not know what influence the indels observed here have had on the biology and evolutionary history of the organisms concerned. While it is likely that the majority of incorporated indels are neutral or near-neutral in nature, we can be certain that some have been subject to selection since indels have been implicated in many human diseases and thus clearly impact protein function (Stenson *et al*, 2003; http://www.hgmd.cf.ac.uk). Indeed some have such high segregating frequencies in certain populations that they must have been selected for, such as the phenylalanine deletion at position 508 in the cystic fibrosis transmembrane conductance regulator protein, which results in cystic fibrosis when homozgyous (Riordan *et al*, 1989), but is believed to be advantageous in heterozygous individuals through providing some degree of resistance to infections disease(s) (Poolman & Galvani, 2007).

Surface loops in proteins are known to be less constrained and more accepting of indels (Pascarella & Argos, 1992; Taylor *et al*, 2004; Reeves *et al*, 2006; de la Chaux *et al*, 2007; Jiang & Blouin, 2007; Guo *et al*, 2012; Tóth-Petróczy & Tawfik, 2013). Thus such regions, particularly if they incorporate insertions providing more template for selection to act upon, may provide fertile grounds

for evolutionary innovation. As better quality data become available through further improvements in sequencing technology and alignment algorithms, we will be able to gain further insight into the formation of indel events and a better understanding of the influence that they have had on protein evolution.

3.3 Asymmetric evolution following gene duplication

There is little doubt that gene duplication is the predominant source of new genes in eukaryotic genomes. However, we still know little in detail about the mechanism and frequency of events that lead to duplication, and still less about the subsequent evolutionary history and dynamics following duplication (Lynch & Katju, 2004; Conant & Wolfe, 2008; Innan & Kondrashov 2010). While there is now reasonable evidence to suggest that there were two rounds of whole genome evolution at the base of the vertebrate lineage (Dehal & Boore, 2005; Putnam *et al*, 2008), and another in teleost fishes (Jaillon *et al*, 2004), with the exception of *Xenopus laevis*, which has been observed in numerous forms of polyploidy (Kobel & Du Pasquier, 1979), no other cases have been described in higher animals. Nevertheless, viable duplications at smaller scales, ranging from individual base pairs to copy-number variants (CNVs) of megabases in length, continue to occur and segregate in animal populations (Redon *et al*, 2006; Zhang *et al*, 2009). CNVs can form by non-allelic homologous recombination, or through tandem duplication resulting in variable number tandem repeat regions, with the former tending to be responsible for longer events and the latter for shorter events (Conrad *et al*, 2010). When CNVs and genes overlap, there is the potential for the establishment of a hotbed of new

innovation – not only may having two or more identical copies of a particular gene be directly advantageous (Perry *et al*, 2007; Konrad *et al*, 2011), but having an extra redundant copy may allow evolution to explore the adaptive landscape leading to increased efficiency, or further, to permit one of the duplicate copies to explore a completely new role in a neofunctional manner (Force *et al*, 1999).

Here we chose to look at individual gene duplicate pairs that we assume to be fixed within the corresponding genomes. We chose rodents because they are fast-evolving, and therefore the sample size of such pairs was relatively high; human and chimpanzee diverged too recently to provide a decent sample size, while comparing human to macaque would likely have proven problematic due to the dubious quality of the current macaque genome assembly (Han *et al*, 2009, Laurie *et al*, 2012). Nevertheless we still ran into some problems with the rat genome, having to exclude a number of putative rat-specific duplicates because we could not find evidence that the genes are expressed. This does not mean that the genes removed may not be *bona fide*, but we prefer to err on the side of caution. We chose to focus on individual duplicates because the evolutionary dynamics associated with being the member of a family where multiple duplication events have occurred are likely to be markedly different, and issues of repetitive events and gene conversion resulting in concerted evolution become a concern (Ezawa *et al*, 2006). While gene conversion may have had some historic influence on a fraction of our dataset, since it most commonly affects closely-spaced tandem duplicates and tends to result in a reduction in sequence divergence, the results reported here can be regarded as conservative with regards to the degree of evolution and asymmetry observed.

We used Ensembl to identify our initial gene list and then built our own trees using a distinct maximum likelihood method (Felsenstein, 2005) to check for consistency with those predicted by Compara (Vilella *et al*, 2009). This was particularly important since we wanted to compare genes that had duplicated before and after separation of the mouse and rat lineages, and having confidence in the timing of the duplication with respect to speciation was essential. For the set that had duplicated prior to the rat and mouse lineages diverging, the timing of the speciation event, approximately 17 million years ago (Douzery *et al*, 2003; Gibbs *et al*, 2004), provides an additional historic time point allowing us to compare evolutionary rates before and after this moment. We used dS, which increases in a manner approximately proportional to time, to split the dataset into two further pairs of time periods for each of the datasets. This structure allowed us to compare rates of evolution shortly after duplication and later, following the passage of time, in each dataset. Comparison with rates in the ancestral pre-duplication branch in the pre-speciation duplication dataset was facilitated by the use of single-copy orthologs from two outgroups – firstly human since it provides the most reliable sequence, and then a further mammalian species from the Laurasiatherian superorder, utilising cow when possible since our previous experience has shown it to be a relatively high quality genome assembly. In both datasets we observed a marked increase in evolutionary rate, measured as dN/dS, immediately following duplication, which gradually returns to pre-duplication rates, presumably as genes become stabilised in their new roles. Importantly, however, this increase was restricted to just one member of each pair, while the other maintained pre-duplication levels.

Using the branch-site test, we found a slightly higher overall frequency of branches testing positive for selection relative to another study which reported that approximately 10% of rat and mouse specific duplications, when including cases of large gene families, had evidence of positive selection (Han *et al*, 2009). Of note however, we observe a much higher frequency of branches testing positive in our older pre-speciation duplication dataset. This may be because the branch-site model may have more power to discriminate selection in genes of this age. For example, Gharib & Robinson-Rechavi (2013) have shown in simulations that the branch-site test has greatest sensitivity with ranges of dS between 0.1 and 0.4. The mean dS for the pre-speciation duplication dataset was between 0.15 and 0.24, whereas that for the post-speciation duplication set ranged from 0.05 to 0.09, and thus there may be some false-negatives in the latter dataset.

It should be noted that we are reporting on extant functional genes, and not observing all of the duplications that have occurred in these lineages, since any that have subsequently resulted in pseudogenisation will not have been included here[11]. Thus we cannot estimate the overall frequency of gene duplication, nor the probability that duplicates will become fixed or pseudogenised post-fixation. While a substantial proportion of our duplicates appear to have formed through retroduplication, in contrast to the observations of Cusack and Wolfe (2006), removal of retrogenes did not affect our asymmetry results i.e. we still observed asymmetry in evolutionary rate in the genes that had undergone DNA-based duplication.

11 Of note, a number of the rat genes which we discarded since we couldn't find evidence of expression were tagged as “known pseudogene” by Ensembl. Our requirement that we have some evidence of expression from EST data is a step that has often been omitted in similar studies, thus likely biasing previously reported results.

It would appear that some of the fast-evolving proteins observed here are experiencing relaxation of selective constraint, while others are undergoing positive selection. It seems likely that the majority of young duplicates initially evolve under relaxation of selective constraint and that only a few reach a point whereby they come under selective constraint, with the remainder degenerating into pseudogenes. We see little evidence for subfunctionalisation here in agreement with previous studies (e.g. Huminiecki & Wolfe, 2004). There are a number of reasons why this may be the case. Firstly, the classical subfunctionalisation model (Hughes, 1994) requires that the parent gene has at least two distinct functions to be subfunctionalised, which rules out many genes, whereas neofunctionalisation has no such requirement. It has also been shown that subfunctionalisation is more likely to occur at smaller population sizes, perhaps consisting of less than 5,000-10,000 individuals (Lynch *et al*, 2001), which is likely to be significantly smaller than the effective population size of the rodents under study here. Furthermore subfunctionalisation may be more likely to occur through means of regulatory sequence change (e.g. Kleinjan *et al*, 2008, Farré & Albà, 2010), which we have not attempted to investigate here. Finally, the act of duplication itself may lead to immediate neofunctionalisation if, as in the case of retrogenes, the new copy forms at a distance from its original regulatory context, but it is difficult to envisage how subfunctionalisation could arise in an immediate manner. However, evidence for subfunctionalisation has been found in organisms that have undergone whole genome duplication (Postlethwait *et al*, 2004; Hellsten *et al*, 2007; Rutter *et al*, 2012), and perhaps may play a more important role in this scenario.

Previous studies that have investigated the fate of duplicate genes from a functional perspective using microarray expression data have found mixed results. From this perspective neofunctionalisation is envisaged as expression in a novel tissue relative to the progenitor copy, whereas subfunctionalisation is division of expression between tissues. Huminiecki & Wolfe (2004), utilising data from the gene expression atlas (Su *et al*, 2002), found that duplicated genes from human and mouse (they included examples from large gene families) show narrower expression across the sixteen tissues analysed, than did one-to-one orthologs, and interestingly, the larger the gene family, the lower the expression breadth and the greater the tissue specificity of individual genes. Analysing expression in a wider set of tissues for primate and human-specific duplicates, Farré and Albà (2010) found the same trend and also noted a significant decrease in expression intensity in duplicate genes. While both these papers report evidence for neofunctionalisation at the expression level, they also make the point that this will again be easier to observe (i.e. expression in a tissue for which the orthologous copy is not observed) than will subfunctionalisation of expression. Indeed they each report possible examples of subfunctionalisation in progress, though there was only one clear cut case where duplicates are expressed in completely distinct tissue sets relative to the presumed ancestral distribution (Huminiecki & Wolfe, 2004). These observations may be further complicated by the observation that most duplicated genes have likely changed their expression pattern with respect to the ancestral state (Huminiecki & Wolfe, 2004), which is typically inferred from the pattern observed for an ortholog in an outgroup species. Here we exploited recent expression data derived from RNAseq analyses (Brawand *et al*, 2011) to investigate expression divergence in part of our dataset. In theory

RNAseq should be able to provide more sensitive measurement of gene expression, with less noise than traditional microarray analyses. Unfortunately there were a relatively small number of cases where we had expression characterised for all three genes in a trio (i.e. mouse duplicate pair and human ortholog). In spite of this we found a correlation between divergence in tissue expression and evolutionary rates, and a few cases that were indicative of neofunctionalisation in terms of expression in a unique tissue in one of the daughter copies. As more complete expression data of this type becomes available we will be able to see whether this is a general pattern.

The studies mentioned above are weakened by the fact that we only have tissue expression data from a subset of all possible tissues, and thus we cannot really be sure of the true expression breadth of the genes concerned. Furthermore, in every case a cut-off point determining what is to be considered biologically significant expression has to be made, and the validity of this cut-off can be questioned. A further issue is that any punctual measure of expression, at best only reports what is happening in a particular class of cell at a particular point in time. As many genes are only transitorily expressed, as and when required by the cell, they will not necessarily be observed in analyses of this type. Furthermore it has been argued that simple classification as subfunctionalisation or neofunctionalisation will often be too narrow (Huminiecki & Wolfe, 2004; He & Zhang, 2005, Rastogi & Liberles, 2005), and that it is often a matter of perspective; paralogs may become subfunctionalised with respect to their ancestral pattern of expression, but may also become neofunctionalised if the pattern of expression extends to tissues where the progenitor was not previously expressed.

Thus discussion as to what is the predominant mode of evolution following duplication, be it subfunctionalisation, neofunctionalisation, or a mixture of both remains undetermined. Here we observed pronounced sequence evolution, sometimes accompanied by evidence of positive selection, generally affecting just one of the copies of a gene that has undergone a single round of duplication in a large number of cases. Furthermore, in a fraction of cases we observed tissue-specific expression in one of the duplicate copies. This supports the hypothesis that one copy, usually the parent which retains its original genomic context, maintains its prior-to-duplication role while the new daughter copy is freed from selective constraint, and in cases where it does not undergo pseudogenisation, can evolve to take on a new role fitting the neofunctionalisation model of Ohno (1970) and Force *et al* (1999). However, evidence from other work shows that subfunctionalisation does also occur, and it is likely that organisms with different life-histories may show different biases towards one mode of evolution or the other following gene duplication.

CONCLUSIONS

- Choice of alignment program can drastically affect downstream comparative genomics analyses, due to the artefact of over-alignment. PRANK performs better than other popular algorithms in this dataset, particularly with regards to the identification of insertions.

- Quality of genome assemblies also affect analyses of this type, and hence some form of quality control is desirable. Here we found the macaque assembly to be of dubious quality, in spite of having in excess of five-fold coverage. Future improvements in sequencing quality and gene prediction algorithms will hopefully abolish this requirement.

- Indel frequency is elevated in low-complexity regions of proteins, and in regions of amino acid tandem repeats. Insertions in particular are more commonly observed in these regions.

- Insertions appear to be more strongly selected against than deletions in these mammalian proteins.

- Indel frequency and evolutionary rate are associated at the level of the protein. It remains unclear if this is due to a direct causal relationship or to a general reduction in selective constraint upon the underlying sequence.

- We do not observe a universal deletion bias in these lineages, and propose that previous reports may be to some extent a result of artefacts in the alignment process.

- We do observe a distinct deletion bias in the ancestral rodent branch.

- We observe a marked increase in evolutionary rate immediately following gene duplication. This increase is restricted to just one of the duplicate copies, and over time it returns to pre-duplication levels.

- In approximately 15-30% of cases, increase in evolutionary rate following gene duplication is accompanied by evidence for positive selection as detected using the branch-site model.

- Divergence in tissue expression is elevated following gene duplication, and we find tentative evidence to support neofunctionalisation following duplication in our dataset, but no evidence of subfunctionalisation.

ANNEX

List of papers associated with this thesis.

Journal Articles

Pegueroles C, **Laurie S**, Albà MM (2013). Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Molecular Biology and Evolution*. Advance access online 26/4/13.

Villanueva-Cañas JL, **Laurie S**, Albà MM (2013). Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biology and Evolution* 5:457-467.

Laurie S, Toll-Riera M, Radó-Trilla N, Albà MM (2012). Sequence shortening in the rodent ancestor. *Genome Research* 22:478-485.

Toll-Riera M, **Laurie S**, Albà MM (2011). Lineage-specific variation in intensity of natural selection in mammals. *Molecular Biology and Evolution* 28:383-398.

Book Chapter

Toll-Riera M, **Laurie S**, Radó-Trilla N, Albà MM (2011). Partial gene duplication and the formation of novel genes. In : F. Friedberg (ed.) *Gene Duplication*. InTech.

Poster Presentations

Annual Meeting of the Society for Molecular Biology and Evolution, Dublin, 2012.
Title: *Is there really a universal deletion bias?*

XI Jornadas de Bioinformática, Barcelona, 2012.
Title: *Sequence shortening in the rodent ancestor*

Annual Meeting of the Society for Molecular Biology and Evolution, Lyon, 2010.
Title: *Characterisation of coding sequence indels in mammalian evolution*

REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., et al. (2000). The genome sequence of *Drosophila melanogaster*. Science (New York, N.Y.), 287(5461), 2185–95.

Albà, M. M., & Guigó, R. (2004). Comparative analysis of amino acid repeats in rodents and humans. Genome research, 14(4), 549–54. doi:10.1101/gr.1925704

Bakewell, M. A., Shi, P., & Zhang, J. (2007). More genes underwent positive selection in chimpanzee evolution than in human evolution. Proceedings of the National Academy of Sciences of the United States of America, 104(18), 7489–94. doi:10.1073/pnas.0701705104

Bedford, T., & Hartl, D. L. (2008). Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in *Drosophila*. Molecular biology and evolution, 25(8), 1631–8. doi:10.1093/molbev/msn112

Betrán, E., Thornton, K., & Long, M. (2002). Retroposed new genes out of the X in *Drosophila.* Genome research, 12(12), 1854–9. doi:10.1101/gr.6049

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise. Genome research, 14(5), 988–95. doi:10.1101/gr.1865504

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome research, 14(4), 708–15. doi:10.1101/gr.1933104

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., et al. (2011). The evolution of gene expression levels in mammalian organs. Nature, 478(7369), 343–8. doi:10.1038/nature10532

Bridges, C. B. (1936). The bar “gene” a duplication. Science (New York, N.Y.), 83(2148), 210–1. doi:10.1126/science.83.2148.210

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Journal of molecular biology, 268(1), 78–94. doi:10.1006/jmbi.1997.0951

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution, 17(4), 540–52.

Chamary, J. V, Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature reviews. Genetics, 7(2), 98–108. doi:10.1038/nrg1770

Chen, F.-C., Chen, C.-J., & Chuang, T.-J. (2007). INDELSCAN: a web server for comparative identification of species-specific and non-species-specific insertion/deletion events. Nucleic acids research, 35(Web Server issue), W633–8. doi:10.1093/nar/gkm350

Chen, F.-C., Chen, C.-J., Li, W.-H., & Chuang, T.-J. (2010). Gene family size conservation is a good indicator of evolutionary rates. Molecular biology and evolution, 27(8), 1750–8. doi:10.1093/molbev/msq055

Chen, J.-Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., & Tian, D. (2009). Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. Molecular biology and evolution, 26(7), 1523–31. doi:10.1093/molbev/msp063

Church, D. M., Goodstadt, L., Hillier, L. W., Zody, M. C., Goldstein, S., She, X., Bult, C. J., et al. (2009). Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS biology, 7(5), e1000112. doi:10.1371/journal.pbio.1000112

Clark, A. G. (1994). Invasion and maintenance of a gene duplication. Proceedings of the National Academy of Sciences of the United States of America, 91(8), 2950–4.

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nature reviews. Genetics, 9(12), 938–50. doi:10.1038/nrg2482

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature, 464(7289), 704–12. doi:10.1038/nature08516

Cooper, G. M., Brudno, M., Stone, E. A., Dubchak, I., Batzoglou, S., & Sidow, A. (2004). Characterization of evolutionary rates and constraints in three Mammalian genomes. Genome research, 14(4), 539–48. doi:10.1101/gr.2034704

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J., & Clamp, M. (2004). The Ensembl automatic gene annotation system. Genome research, 14(5), 942–50. doi:10.1101/gr.1858004

Cusack, B. P., & Wolfe, K. H. (2007). Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Molecular biology and evolution, 24(3), 679–86. doi:10.1093/molbev/msl199

Darwin, C. (1859). On the origin of species by means of natural selection. John Murray.

Davis, J. C., & Petrov, D. A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. PLoS biology, 2(3), E55. doi:10.1371/journal.pbio.0020055

De Jong, W. W., & Rydén, L. (1981). Causes of more frequent deletions than insertions in mutations and protein evolution. Nature, 290(5802), 157–9.

de la Chaux, N., Messer, P. W., & Arndt, P. F. (2007). DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. BMC evolutionary biology, 7, 191. doi:10.1186/1471-2148-7-191

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS biology, 3(10), e314. doi:10.1371/journal.pbio.0030314

Deininger, P. L., & Batzer, M. A. (2002). Mammalian retroelements. Genome research, 12(10), 1455–65. doi:10.1101/gr.282402

Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N., & Hahn, M. W. (2006). The evolution of mammalian gene families. PloS one, 1, e85. doi:10.1371/journal.pone.0000085

Des Marais, D. L., & Rausher, M. D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature, 454(7205), 762–5. doi:10.1038/nature07092

Dessimoz, C., & Gil, M. (2010). Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome biology, 11(4), R37. doi:10.1186/gb-2010-11-4-r37

Dong, D., He, G., Zhang, S., & Zhang, Z. (2009). Evolution of olfactory receptor genes in primates dominated by birth-and-death process. Genome biology and evolution, 1, 258–64. doi:10.1093/gbe/evp026

Douzery, E. J. P., Delsuc, F., Stanhope, M. J., & Huchon, D. (2003). Local molecular clocks in three nuclear genes: divergence times for rodents and other mammals and incompatibility among fossil calibrations. Journal of molecular evolution, 57 Suppl 1, S201–13. doi:10.1007/s00239-003-0028-x

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57–74. doi:10.1038/nature11247

Dunning Hotopp, J. C. (2011). Horizontal gene transfer between bacteria and animals. Trends in genetics : TIG, 27(4), 157–63. doi:10.1016/j.tig.2011.01.005

Eddy, S. R. (2004). What is a hidden Markov model? Nature biotechnology, 22(10), 1315–6. doi:10.1038/nbt1004-1315

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), 1792–7. doi:10.1093/nar/gkh340

Ellegren, H., Smith, N. G. C., & Webster, M. T. (2003). Mutation rate variation in the mammalian genome. Current opinion in genetics & development, 13(6), 562–8.

Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstock, G. M., Adelson, D. L., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science (New York, N.Y.), 324(5926), 522–8. doi:10.1126/science.1169588

Emerson, J. J., Kaessmann, H., Betrán, E., & Long, M. (2004). Extensive gene traffic on the mammalian X chromosome. Science (New York, N.Y.), 303(5657), 537–40. doi:10.1126/science.1090042

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., et al. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS one, 7(11), e47768. doi:10.1371/journal.pone.0047768

Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E., Blake, J. A., Anagnostopoulos, A., Baldarelli, R. M., et al. (2005). The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. Nucleic acids research, 33(Database issue), D471–5. doi:10.1093/nar/gki113

Ezawa, K., Ikeo, K., Gojobori, T., & Saitou, N. (2011). Evolutionary patterns of recently emerged animal duplogs. Genome biology and evolution, 3, 1119–35. doi:10.1093/gbe/evr074

Ezawa, K., Oota, S., & Saitou, N. (2006). Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Genome-wide search of gene conversions in duplicated genes of mouse and rat. Molecular biology and evolution, 23(5), 927–40. doi:10.1093/molbev/msj093

Fan, Y., Wang, W., Ma, G., Liang, L., Shi, Q., & Tao, S. (2007). Patterns of insertion and deletion in Mammalian genomes. Current genomics, 8(6), 370–8. doi:10.2174/138920207783406479

Farré, D., & Albà, M. M. (2010). Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. Molecular biology and evolution, 27(2), 325–35. doi:10.1093/molbev/msp242

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6.

Feng, D. F., & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of molecular evolution, 25(4), 351–60.

Fletcher, W., & Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Molecular biology and evolution, 27(10), 2257–67. doi:10.1093/molbev/msq115

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., et al. (2013). Ensembl 2013. Nucleic acids research, 41(Database issue), D48–55. Doi:10.1093/nar/gks1236

Florea, L., Souvorov, A., Kalbfleisch, T. S., & Salzberg, S. L. (2011). Genome assembly has a major impact on gene content: a comparison of annotation in two *Bos taurus* assemblies. PloS one, 6(6), e21400. doi:10.1371/journal.pone.0021400

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

Gaffney, D. J., & Keightley, P. D. (2006). Genomic selective constraints in murid noncoding DNA. PLoS genetics, 2(11), e204. doi:10.1371/journal.pgen.0020204

Gharib, W. H., & Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. Molecular biology and evolution. doi:10.1093/molbev/mst062

Gibbs, R. A, Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature, 428(6982), 493–521. doi:10.1038/nature02426

Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., Remington, K. A., et al. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. Science (New York, N.Y.), 316(5822), 222–34. doi:10.1126/science.1139247

Gilad, Y., Przeworski, M., & Lancet, D. (2004). Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. PLoS biology, 2(1), E5. Doi:10.1371/journal.pbio.0020005

Golubchik, T., Wise, M. J., Easteal, S., & Jermiin, L. S. (2007). Mind the gaps: evidence of bias in estimates of multiple sequence alignments. Molecular biology and evolution, 24(11), 2433–42. doi:10.1093/molbev/msm176

Graur, D., Shuali, Y., & Li, W. H. (1989). Deletions in processed pseudogenes accumulate faster in rodents than in humans. Journal of molecular evolution, 28(4), 279–85.

Gregory, T. R. (2003). Is small indel bias a determinant of genome size? Trends in genetics : TIG, 19(9), 485–8. doi:10.1016/S0168-9525(03)00192-6

Gregory, T. R. (2004). Insertion-deletion biases and the evolution of genome size. Gene, 324, 15–34.

Gross, S. S., Do, C. B., Sirota, M., & Batzoglou, S. (2007). CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. Genome biology, 8(12), R269. doi:10.1186/gb-2007-8-12-r269

Gu, X., & Li, W. H. (1992). Higher rates of amino acid substitution in rodents than in humans. Molecular phylogenetics and evolution, 1(3), 211–4.

Gu, Z., Nicolae, D., Lu, H. H.-S., & Li, W. H. (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. Trends in genetics : TIG, 18(12), 609–13.

Guo, B., Zou, M., & Wagner, A. (2012). Pervasive indels and their evolutionary dynamics after the fish-specific genome duplication. Molecular biology and evolution, 29(10), 3005–22. doi:10.1093/molbev/mss108

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. The Journal of heredity, 100(5), 605–17. doi:10.1093/jhered/esp047

Hahn, M. W., Demuth, J. P., & Han, S.-G. (2007). Accelerated rate of gene gain and loss in primates. Genetics, 177(3), 1941–9. doi:10.1534/genetics.107.080077

Han, M. V, Demuth, J. P., McGrath, C. L., Casola, C., & Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. Genome research, 19(5), 859–67. doi:10.1101/gr.085951.108

Hardison, R. C., Roskin, K. M., Yang, S., Diekhans, M., Kent, W. J., Weber, R., Elnitski, L., et al. (2003). Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome research, 13(1), 13–26. doi:10.1101/gr.844103

He, X., & Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics, 169(2), 1157–64. doi:10.1534/genetics.104.037051

He, X., & Zhang, J. (2006). Higher duplicability of less important genes in yeast genomes. Molecular biology and evolution, 23(1), 144–51. doi:10.1093/molbev/msj015

Hellsten, U., Khokha, M. K., Grammer, T. C., Harland, R. M., Richardson, P., & Rokhsar, D. S. (2007). Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. BMC biology, 5, 31. doi:10.1186/1741-7007-5-31

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America, 89(22),10915–9.

Higgins, D. G., Blackshields, G., & Wallace, I. M. (2005). Mind the gaps: progress in progressive alignment. Proceedings of the National Academy of Sciences of the United States of America, 102(30), 10411–2. doi:10.1073/pnas.0504801102

Higgins, D. G., & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, 73(1), 237–44.

Hinkley, C. S., & Ismaili, L. (2012). A rapid genotyping assay for segregating human olfactory receptor pseudogenes. Journal of biomolecular techniques : JBT, 23(3), 84–9. doi:10.7171/jbt.12-2303-001

Hodgkinson, A., & Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. Nature reviews. Genetics, 12(11), 756–66. doi:10.1038/nrg3098

Hubisz, M. J., Lin, M. F., Kellis, M., & Siepel, A. (2011). Error and error mitigation in low-coverage genome assemblies. PloS one, 6(2), e17034. doi:10.1371/journal.pone.0017034

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. Proceedings. Biological sciences / The Royal Society, 256(1346), 119–24. doi:10.1098/rspb.1994.0058

Hughes, A. L., & Nei, M. (1989). Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proceedings of the National Academy of Sciences of the United States of America, 86(3), 958–62.

Huminiecki, L., & Wolfe, K. H. (2004). Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome research, 14(10A), 1870–9. doi:10.1101/gr.2705204

Imamura, H., Karro, J. E., & Chuang, J. H. (2009). Weak preservation of local neutral substitution rates across mammalian genomes. BMC evolutionary biology, 9, 89. doi:10.1186/1471-2148-9-89

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nature reviews. Genetics, 11(2), 97–108. doi:10.1038/nrg2689

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. Nature, 431(7011), 931–45. doi:10.1038/nature03001

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature, 431(7011), 946–57. doi:10.1038/nature03025

Jiang, H., & Blouin, C. (2007). Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. BMC bioinformatics, 8, 444. doi:10.1186/1471-2105-8-444

Jordan, G., & Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Molecular biology and evolution, 29(4), 1125–39. doi:10.1093/molbev/msr272

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research, 110(1-4), 462–7. doi:10.1159/000084979

Kaessmann, H., Vinckenbosch, N., & Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. Nature reviews. Genetics, 10(1), 19–31. doi:10.1038/nrg2487

Kamal, M., Xie, X., & Lander, E. S. (2006). A large family of ancient repeat elements in the human genome is under strong selection. Proceedings of the National Academy of Sciences of the United States of America, 103(8), 2740–5. doi:10.1073/pnas.0511238103

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, 30(14), 3059–66.

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature, 217(5129), 624–6.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Cambridge University Press.

King, J. L., & Jukes, T. H. (1969). Non-Darwinian evolution. Science (New York, N.Y.), 164(3881), 788–98.

Kleinjan, D. A., Bancewicz, R. M., Gautier, P., Dahm, R., Schonthaler, H. B., Damante, G., Seawright, A., et al. (2008). Subfunctionalization of duplicated zebrafish pax6 genes by cis-regulatory divergence. PLoS genetics, 4(2), e29. doi:10.1371/journal.pgen.0040029

Knowles, D. G., & McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. Genome research, 19(10), 1752–9. doi:10.1101/gr.095026.109

Kobel, H. R., & Du Pasquier, L. (1979). Hyperdiploid species hybrids for gene mapping in *Xenopus*. Nature, 279(5709), 157–8.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Selection in the evolution of gene duplications. Genome biology, 3(2), RESEARCH0008.

Konrad, A., Teufel, A. I., Grahnen, J. A., & Liberles, D. A. (2011). Toward a general model for the evolutionary dynamics of gene duplicates. Genome biology and evolution, 3, 1197–209. doi:10.1093/gbe/evr093

Kuo, C.-H., & Ochman, H. (2009). Deletional bias across the three domains of life. Genome biology and evolution, 1, 145–52. doi:10.1093/gbe/evp016

Kvikstad, E. M., Chiaromonte, F., & Makova, K. D. (2009). Ride the wavelet: A multiscale analysis of genomic contexts flanking small insertions and deletions. Genome research, 19(7), 1153–64. doi:10.1101/gr.088922.108

Kvikstad, E. M., Tyekucheva, S., Chiaromonte, F., & Makova, K. D. (2007). A macaque's-eye view of human insertions and deletions: differences in mechanisms. PLoS computational biology, 3(9), 1772–82. doi:10.1371/journal.pcbi.0030176

Labbé, P., Berthomieu, A., Berticat, C., Alout, H., Raymond, M., Lenormand, T., & Weill, M. (2007). Independent duplications of the acetylcholinesterase gene conferring insecticide resistance in the mosquito *Culex pipiens*. Molecular biology and evolution, 24(4), 1056–67. doi:10.1093/molbev/msm025

Landan, G., & Graur, D. (2007). Heads or tails: a simple reliability check for multiple sequence alignments. Molecular biology and evolution, 24(6), 1380–3. doi:10.1093/molbev/msm060

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., et al. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860–921. doi:10.1038/35057062

Leushkin, E. V, Bazykin, G. A., & Kondrashov, A. S. (2012). Insertions and deletions trigger adaptive walks in *Drosophila* proteins. Proceedings. Biological sciences / The Royal Society, 279(1740), 3075–82. doi:10.1098/rspb.2011.2571

Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Molecular biology and evolution, 4(3), 203–21.

Li, L., Huang, Y., Xia, X., & Sun, Z. (2006). Preferential duplication in the sparse part of yeast protein interaction network. Molecular biology and evolution, 23(12), 2467–73. doi:10.1093/molbev/msl121

Li, W.-H. (1997). Molecular evolution. Sinauer Associates Incorporated.

Liang, H., & Li, W.-H. (2007). Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends in genetics : TIG, 23(8), 375–8. doi:10.1016/j.tig.2007.04.005

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature, 478(7370), 476–82. doi:10.1038/nature10530

Löytynoja, A., & Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America, 102(30), 10557–62. doi:10.1073/pnas.0409137102

Löytynoja, A., & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science (New York, N.Y.), 320(5883), 1632–5. doi:10.1126/science.1158395

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

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics, 154(1), 459–73.

Lynch, M., O'Hely, M., Walsh, B., & Force, A. (2001). The probability of preservation of a newly arisen gene duplicate. Genetics, 159(4), 1789–804.

Lynch, M., & Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. Trends in genetics : TIG, 20(11), 544–9. doi:10.1016/j.tig.2004.09.001

Makalowski, W. (2001). Are we polyploids? A brief history of one hypothesis. Genome research, 11(5), 667–70. doi:10.1101/gr.188801

Makova, K. D., & Li, W.-H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. Genome research, 13(7), 1638–45. doi:10.1101/gr.1133803

Mallick, S., Gnerre, S., Muller, P., & Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. Genome research, 19(5), 922–33. doi:10.1101/gr.086512.108

Markova-Raina, P., & Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. Genome research, 21(6), 863–74. doi:10.1101/gr.115949.110

McCarrey, J. R., & Thomas, K. (1987). Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature, 326(6112), 501–5. doi:10.1038/326501a0

McDonald, M. J., Wang, W.-C., Huang, H.-D., & Leu, J.-Y. (2011). Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. PLoS biology, 9(6), e1000622. doi:10.1371/journal.pbio.1000622

Messer, P. W., & Arndt, P. F. (2007). The majority of recent short DNA insertions in the human genome are tandem duplications. Molecular biology and evolution, 24(5), 1190–7. doi:10.1093/molbev/msm035

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic acids research, 41(Database issue), D64–9. doi:10.1093/nar/gks1048

Mikkelsen, T. S., Wakefield, M. J., Aken, B., Amemiya, C. T., Chang, J. L., Duke, S., Garber, M., et al. (2007). Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. Nature, 447(7141), 167–77. doi:10.1038/nature05805

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. Genome research, 16(9), 1182–90. doi:10.1101/gr.4565806

Muller, J., Creevey, C. J., Thompson, J. D., Arendt, D., & Bork, P. (2010). AQUA: automated quality improvement for multiple sequence alignments. Bioinformatics (Oxford, England), 26(2), 263–5. doi:10.1093/bioinformatics/btp651

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harbor symposia on quantitative biology, 51 Pt 1, 263–73.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3), 443–53.

Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics, 148(3), 929–36.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology, 302(1), 205–17. doi:10.1006/jmbi.2000.4042

Novoa, E. M., & Ribas de Pouplana, L. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. Trends in genetics : TIG, 28(11), 574–81. doi:10.1016/j.tig.2012.07.006

Nozawa, M., Suzuki, Y., & Nei, M. (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proceedings of the National Academy of Sciences of the United States of America, 106(16), 6700–5. doi:10.1073/pnas.0901855106

Ohno, S. (1970). Evolution by gene duplication (p. 160). Springer-Verlag.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. Nature, 246(5428), 96–8.

Oldmeadow, C., Mengersen, K., Mattick, J. S., & Keith, J. M. (2010). Multiple evolutionary rate classes in animal genome evolution. Molecular biology and evolution, 27(4), 942–53. doi:10.1093/molbev/msp299

Ophir, R., & Graur, D. (1997). Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene, 205(1-2), 191–202.

Pascarella, S., & Argos, P. (1992). Analysis of insertions/deletions in protein structures. Journal of molecular biology, 224(2), 461–71.

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., & Pupko, T. (2010). GUIDANCE: a web server for assessing alignment confidence scores. Nucleic acids research, 38(Web Server issue), W23–8. doi:10.1093/nar/gkq443

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nature genetics, 39(10), 1256–60. doi:10.1038/ng2123

Petrov, D. A. (2002). Mutational equilibrium model of genome size evolution. Theoretical population biology, 61(4), 531–44.

Ponting, C. P., & Hardison, R. C. (2011). What fraction of the human genome is functional? Genome research, 21(11), 1769–76. doi:10.1101/gr.116814.110

Poolman, E. M., & Galvani, A. P. (2007). Evaluating candidate agents of selective pressure for cystic fibrosis. Journal of the Royal Society, Interface / the Royal Society, 4(12), 91–8. doi:10.1098/rsif.2006.0154

Postlethwait, J., Amores, A., Cresko, W., Singer, A., & Yan, Y.-L. (2004). Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends in genetics : TIG, 20(10), 481–90. doi:10.1016/j.tig.2004.08.001

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. Nature, 453(7198), 1064–71. doi:10.1038/nature06967

Rastogi, S., & Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC evolutionary biology, 5, 28. doi:10.1186/1471-2148-5-28

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., et al. (2006). Global variation in copy number in the human genome. Nature, 444(7118), 444–54. doi:10.1038/nature05329

Reeves, G. A., Dallman, T. J., Redfern, O. C., Akpor, A., & Orengo, C. A. (2006). Structural diversity of domain superfamilies in the CATH database. Journal of molecular biology, 360(3), 725–41. doi:10.1016/j.jmb.2006.05.035

Rieseberg, L. H., & Willis, J. H. (2007). Plant speciation. Science (New York, N.Y.), 317(5840), 910–4. doi:10.1126/science.1137729

Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science (New York, N.Y.), 245(4922), 1066–73.

Robinson-Rechavi, M., & Laudet, V. (2001). Evolutionary rates of duplicate genes in fish and mammals. Molecular biology and evolution, 18(4), 681–3.

Rutter, M. T., Cross, K. V, & Van Woert, P. A. (2012). Birth, death and subfunctionalization in the *Arabidopsis* genome. Trends in plant science, 17(4), 204–12. doi:10.1016/j.tplants.2012.01.006

Scannell, D. R., & Wolfe, K. H. (2008). A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome research, 18(1), 137–47. doi:10.1101/gr.6341207

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., & Graur, D. (2009). Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome biology and evolution, 1, 114–8. doi:10.1093/gbe/evp012

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., et al. (2003). Human-mouse alignments with BLASTZ. Genome research, 13(1), 103–7. doi:10.1101/gr.809403

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Current opinion in genetics & development, 9(6), 657–63.

Stenson, P. D., Ball, E. V, Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., et al. (2003). Human Gene Mutation Database (HGMD): 2003 update. Human mutation, 21(6), 577–81. doi:10.1002/humu.10212

Studer, R. A., Penel, S., Duret, L., & Robinson-Rechavi, M. (2008). Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. Genome research, 18(9), 1393–402. doi:10.1101/gr.076992.108

Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., et al. (2002). Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences of the United States of America, 99(7), 4465–70. doi:10.1073/pnas.012025199

Tanay, A., & Siggia, E. D. (2008). Sequence context affects the rate of short insertions and deletions in flies and primates. Genome biology, 9(2), R37. doi:10.1186/gb-2008-9-2-r37

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. Nature reviews. Genetics, 12(10), 692–702. doi:10.1038/nrg3053

Taylor, M. S., Ponting, C. P., & Copley, R. R. (2004). Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. Genome research, 14(4), 555–66. doi:10.1101/gr.1977804

Thompson, J. D., Higgins, D. G., & 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 research, 22(22), 4673–80.

Thompson, J. D., Plewniak, F., & Poch, O. (1999a). BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics (Oxford, England), 15(1), 87–8.

Thompson, J. D., Plewniak, F., & Poch, O. (1999b). A comprehensive comparison of multiple sequence alignment programs. Nucleic acids research, 27(13), 2682–90.

Thompson, J. D., Linard, B., Lecompte, O., & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PloS one, 6(3), e18093. doi:10.1371/journal.pone.0018093

Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., et al. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature, 455(7209), 105–8. doi:10.1038/nature07175

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., & Albà, M. M. (2009). Origin of primate orphan genes: a comparative genomics approach. Molecular biology and evolution, 26(3), 603–12. doi:10.1093/molbev/msn281

Tóth-Petróczy, A., & Tawfik, D. S. (2013). Protein insertions and deletions enabled by neutral roaming in sequence space. Molecular biology and evolution, 30(4), 761–71. doi:10.1093/molbev/mst003

Vamathevan, J. J., Hasan, S., Emes, R. D., Amrine-Madsen, H., Rajagopalan, D., Topp, S. D., Kumar, V., et al. (2008). The role of positive selection in determining the molecular cause of species differences in disease. BMC evolutionary biology, 8, 273. doi:10.1186/1471-2148-8-273

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., et al. (2001). The sequence of the human genome. Science (New York, N.Y.), 291(5507), 1304–51. doi:10.1126/science.1058040

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., & Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome research, 19(2), 327–35. doi:10.1101/gr.073585.107

Vinckenbosch, N., Dupanloup, I., & Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. Proceedings of the National Academy of Sciences of the United States of America, 103(9), 3220–5. doi:10.1073/pnas.0511307103

Wagner, A. (2000). Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proceedings of the National Academy of Sciences of the United States of America, 97(12), 6579–84. doi:10.1073/pnas.110147097

Walsh, B. (2003). Population-genetic models of the fates of duplicate genes. Genetica, 118(2-3), 279–94.

Wang, Z., Martin, J., Abubucker, S., Yin, Y., Gasser, R. B., & Mitreva, M. (2009). Systematic analysis of insertions and deletions specific to nematode proteins and their proposed functional and evolutionary relevance. BMC evolutionary biology, 9, 23. doi:10.1186/1471-2148-9-23

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature, 420(6915), 520–62. doi:10.1038/nature01262

Wilming, L. G., Gilbert, J. G. R., Howe, K., Trevanion, S., Hubbard, T., & Harrow, J. L. (2008). The vertebrate genome annotation (Vega) database. Nucleic acids research, 36(Database issue), D753–60. doi:10.1093/nar/gkm987

Wolf, J. B. W., Künstner, A., Nam, K., Jakobsson, M., & Ellegren, H. (2009). Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. Genome biology and evolution, 1, 308–19. doi:10.1093/gbe/evp030

Wong, K. M., Suchard, M. A., & Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. Science (New York, N.Y.), 319(5862), 473–6. doi:10.1126/science.1151532

Yang, S., Smit, A. F., Schwartz, S., Chiaromonte, F., Roskin, K. M., Haussler, D., Miller, W., et al. (2004). Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. Genome research, 14(4), 517–27. doi:10.1101/gr.1984404

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS, 13(5), 555–6.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular biology and evolution, 15(5), 568–73.

Yang, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. Journal of molecular evolution, 51(5), 423–32.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution, 24(8), 1586–91. doi:10.1093/molbev/msm088

Yang, Z., & Dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. Molecular biology and evolution, 28(3), 1217–28. doi:10.1093/molbev/msq303

Yang, Z., & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular biology and evolution, 19(6), 908–17. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12032247

Yang, Z., Wong, W. S. W., & Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. Molecular biology and evolution, 22(4), 1107–18. doi:10.1093/molbev/msi097

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. Annual review of genomics and human genetics, 10, 451–81. doi:10.1146/annurev.genom.9.081307.164217

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular biology and evolution, 22(12), 2472–9. doi:10.1093/molbev/msi237

Zhang, P., Gu, Z., & Li, W.-H. (2003). Different evolutionary patterns between young duplicate genes in the human genome. Genome biology, 4(9), R56. doi:10.1186/gb-2003-4-9-r56

La impressió d'aquesta tesi ha estat possible gràcies a l'ajut per a la finalització de tesis doctorals de la Fundació IMIM.