Inferring strength of selection in vertebrate genomes

Lél Eöry

Thesis presented for the degree of Doctor of Philosophy

University of Edinburgh

2010

Declaration

Hereby I declare that this thesis was composed by myself, that the work contained herein is my own and no part of this thesis has been submitted to any other university in application for a higher degree.

Lél Eöry

Edinburgh, December 2010.

Acknowledgements

I thank to my supervisor, Brian Charlesworth, for his support and regular discussions during the last year of my PhD. Your comments and suggestions were invaluable and were helping me throughout the writing up of my thesis. I also very much appreciate that you were willing to take me as a PhD student after my third-year of study, if only this change had happened a year earlier. I also thank to Paul Sharp, who became my second supervisor, for his comments and suggestions on different manuscripts. Comments given by Peter Keightley, my former supervisor, on my first year report and on different manuscripts contributed to the quality of my work to a large extent, and he helped me to increase my psychical endurance by providing extreme work conditions during my first and second years.

This project could not get to a happy ending without the encouragement and support of many people, amongst them Markus, Dario, Helen, Sonja, Juan, Rolf, Kati, Dan, Matthew, Silvia, Balazs, Helen, Laura, Gethin, Jayna and Melissa. Special thank goes to Richard Ennos, whose determination to secure a change in my supervision had definitely saved my PhD and helped me to regain the ability to focus on my work.

Finally, I thank to my family here in Edinburgh and in Hungary for their encouragement and support and for the long-term babysittings without which the completion of this PhD would have been impossible. I am greatly indebted to my wife Vera, and our children, Hanga and Abel for their patience and tolerance of the long working hours and lost weekends and also for the strength and hope they filled me up with in times of despair.

Funding for this PhD was provided by the Marie Curie Host Fellowships for Early Stage Training, as part of the 6th Framework Programme of the European Commission. I also acknowledge a Research Associate position with Peter Keightley for a period of five-month.

Table of Contents

Abstract

Protein-coding sequences have long been assumed to evolve under selection, but the quantification of the process at the nucleotide sequence level only started when a simple null model, the neutral theory of molecular evolution, was formulated by Kimura. Several methods were developed, which were based on the assumption that synonymous sites (nucleotides at third codon positions which do not change the encoded amino acid) evolve close to neutrally, and could be used as local neutral standards. Most of our current knowledge on the direction and strength of selection still depends on this simple assumption. One method, notably the non-synonymous to synonymous substitution rate ratio (d_N/d_S) has gained prevalence and is still widely used, in spite of the growing body of evidence that synonymous sites evolve under selection. In this thesis, I quantify the strength of selection in different sequence compartments of mammalian genomes, in order to obtain estimates of their functional importance from comparative genomics analyses. I quantify the fraction of mutations that have been selectively eliminated since the divergence of the species pairs examined, the so called genome wide selective constraint. This in turn is used to approximate the genomic deleterious mutation rate, which is an important parameter for several evolutionary problems. As estimates of selection depend on a large extent on the chosen neutral standard, here I use orthologous transposable elements, so called ancestral repeats, as these have been found to be evolving at a largely neutral fashion, and contain the least number of constrained sites in mammalian genomes. This enables me to quantify the level of selection even at synonymous sites, and the results suggest that these sites indeed evolve under constraint, the consequences of which I discuss. The selective constraint estimates enable me to test some simple hypotheses, such as Ohta's nearly neutral theory of molecular evolution, which suggests that selection is more efficient in species with larger effective population sizes. Beside the choice of neutral standards, there are several additional factors which are known to affect the selective constraint estimates. Here I also test the consequences of one of these, notably when sequences

are not at compositional equilibrium (i.e. their GC content is away from the equilibrium GC content), which predicts that sequences with different GC content should evolve with different rates. This can cause bias in the estimates of level of selection or can even imitate selection in sequences which evolve completely neutrally. This effect is quantified here, and a simple correction is discussed.

Abbreviations

0F: non-synonymous site, zero-fold site

4F: synonymous site, four-fold degenerate site

3'UTR: downstream untranslated region

5'UTR: upstream untranslated region

AR: ancestral repeat

AS: alternative spliced gene

BGC: biased gene conversion

C: selective constraint

d_N/d_S: non-synonymous to synonymous substitution rate ratio

E: expected number of changes

ESE: exon splice enhancer

ESS: exon splice silencer

indel: insertions/deletions

kb: kilobase

L: mutation load

μ mutation rate per generation

M: number of mutations

Mya: million years ago

N_e: long-term effective population size

non-CpG site: nucleotide site which has not been part of an ancestral CpG dinucleotide

O: observed number of changes

p_e: equilibrium GC content

s: selection coefficient

S: A or T to G or C substitution rate

ST: single transcript gene

TE: transposable element

u: G or C to A or T mutation rate

U: deleterious mutation rate

v: A or T to G or C mutation rate

W: G or C to A or T substitution rate

1 Introduction

The most important process in the evolution of DNA sequences is the change in nucleotides with time. Changes in the genetic material, called mutations, result from errors in DNA replication or DNA repair (Arnheim & Calabrese 2009; Li 1997), and lead to variation in the genetic material. The frequency of the resulting new alleles within populations will either increase or decrease depending on two main processes, selection and random genetic drift (Kimura 1962).

The selectionists' view, developed from Darwin's proposal of evolution through natural selection (Darwin 1859), was that the majority of observed differences within or between species are the results of mutations either fixed by adaptive evolution (Fisher 1930; Wright 1931), maintained by mutation-selection balance (Kimura & Crow 1964) or by some form of balancing selection (e.g. Wallace & Dobzhansky 1962). This view started to crumble with the discoveries that 1) amino acid changes are more frequent in less important proteins (Zuckerkandl & Pauling 1965) or in less important regions of proteins (Margoliash & Smith 1965; Zuckerkandl & Pauling 1965); 2) polymorphism was found to be more common in proteins than had previously been expected (Lewontin & Hubby 1966); 3) the genomic rate of nucleotide substitution was estimated by Kimura (1968) to be too high to be compatible with the estimated upper limit of gene substitution by natural selection, predicted by (Haldane 1957) and 4) the frequency of nucleotide change at third positions of codons, which are less likely to change the protein coding sequence, was found to be much higher than at the first two codon positions (King & Jukes 1969). These observations led to the formulation of the neutral theory of molecular evolution by Kimura (1968) and King & Jukes (1969) independently.

1.1 The neutral theory and its consequences

The neutral theory states that most of the new mutations are selectively neutral or have such small effects on fitness that their fate is determined not by selection but by random genetic drift due to gamete sampling in finite populations (Kimura 1983).

The theory does not exclude possible contributions of natural selection to sequence evolution, but contrary to the selectionist theory, which claimed that most new mutations were the results of adaptive evolution, it assumes much lower fractions of mutations under positive selection, as most of these mutations are considered to be already fixed by selection.

1.1.1 The nearly neutral theory

If the selection coefficient (s) – measuring the intensity of natural selection on the genotypes at a locus – is substantial in relation to the effective population size (N_e), i.e. $|s| > 1/N_e$, then the dynamics of the frequency of new alleles depend largely on mutation and selection. Mutations with harmful effects are most likely to be selected against and are to be removed by purifying selection, while those conferring a selective advantage to individuals tend to increase their frequencies through positive selection until reaching fixation, although some advantageous mutations may still become lost due to the effect of random sampling. On the other hand, in the case of strict neutrality, when mutations have no effect on fitness (s=0), their fixation probability depends on random genetic drift alone (Kimura 1983). The fate of those mutations with small effects on fitness ($|s|<1/N_e$) is described by an extension of the neutral theory, referred to as the nearly neutral theory (Ohta 1973; Ohta 1995). These mutations are called effectively neutral if the selective advantage or disadvantage associated with them is so small, relative to N_e, that they behave as if completely neutral. It is important to note, that the notion of effective neutrality depends on N_e. On an absolute scale a mutation with a given s may behave as effectively neutral in a species with small N_e (e.g. in mammals), while selected against in another species with larger N_e (e.g. fruit fly). Many of these mutations have $s \approx 1/N_e$, and if their number is substantial, then a negative correlation is expected between evolutionary

rate and N_e, while no correlation is expected for strictly neutral mutations (Ohta & Gillespie 1996). As the probability of fixation of new mutations depends on their fitness effects, attempts have been made recently to describe the distribution of fitness effects of new mutations (e.g. Eyre-Walker & Keightley 2007; Boyko et al. 2008), but uncertainties still remain and the distribution is largely unknown for most species.

1.1.2 Mutation rate

Since mutations provide the substrate for selection, the rate at which mutations are introduced into genomes play an important role in understanding sequence evolution. Although different types of mutations, e.g. point mutations, insertions and deletions (indels) or translocations (Arnheim & Calabrese 2009), all have an impact on sequence evolution (Wetterbom et al. 2006; Hahn et al. 2007; Eyre-Walker & Keightley 2009), many studies focused solely on the effect of single nucleotide changes (Charlesworth & Charlesworth 1998; Keightley & Eyre-Walker 2000; Eyre-Walker 2006), as these events are the most frequent types of mutations (e.g. CSAC 2005) and reconstructing their evolutionary history has proved to be more straightforward than for the other categories, notably for indels (Cartwright 2009). Here, unless stated otherwise, I refer to single nucleotide changes as mutations, as the analyses presented here focus on the consequences of these on vertebrate genome evolution.

Although it is possible to obtain direct estimates of the point mutation rate per generation, especially for species with short generation times (e.g. Haag-Liautard et al. 2007), estimates are not readily available for mammalian species (but see Kondrashov 2003; Lynch 2010). Nevertheless, as a consequence of the neutral theory, the lack of knowledge of the mutation rate can be overcome by a simple way which provides a means to get estimates of the point mutation rate. In the case of strict neutrality if the rate of point mutations is μ per haploid genome, then at a given nucleotide site, the expected number of mutations in a diploid population per generation is $2N\mu$, where N is the population size. As the fixation probability of a

neutral mutation is 1/(2N) then the observed substitution rate (d) becomes d=2N μ/(2N), i.e. the rate of evolution at neutral sites can be used to estimate the mutation rate (Kimura 1968a).

1.1.3 Predictions of the neutral theory

Another important consequence of the neutral theory is that it has led to many testable predictions on the rates and patterns of nucleotide polymorphisms and substitutions. One of these predictions is that the rate of evolution or level of polymorphism should be lower at sites with higher functional significance. At synonymous sites of codons for example, where mutations do not change the protein coding sequence, the neutral theory correctly predicts higher rates of evolution than at non-synonymous sites, where mutations, by inflicting a change in the amino acid sequence, can frequently lead to a reduction in fitness. This observation contradicted the selectionists' expectation of higher rate of evolution at non-synonymous site due to adaptive evolution, and provided strong evidence for the neutral theory (King & Jukes 1969).

If μ is known, then at any given nucleotide sites, d can be given as d= $\mu Q(S)$ (e.g. Charlesworth & Eyre-Walke 2007), where Q is a function of s, taken from the distribution of fitness effects, and N_e. The equation encapsulates the three factors which jointly drive genome evolution, namely mutation, selection and drift. Based on this equation it follows that there should be variation in divergence or in the level of polymorphism among different nucleotide sequences, due to differences in the distribution of functional sites and their functional importance. Indeed, substitution rates vary widely between pseudogenes, intergenic regions and other sequence types associated with protein coding genes (i.e. synonymous and non-synonymous sites, introns, upstream and downstream untranslated regions) in fruit flies (Haddrill et al. 2005), rodents (Gaffney & Keightley 2005) and hominids (Makalowski & Boguski 1998; Bustamante et al. 2002; Keightley et al. 2005).

Based on sequence data and on an assumed neutral standard the neutral theory provides the null hypothesis for many different tests aim to infer selection on the

sequences. These tests are designed to capture signatures of selection and are frequently based on comparisons of within or between species polymorphism and/or divergence (e.g. the relative rate test – Miyata & Yasunaga (1980), Tajima's D – Tajima (1989), Hudson-Kreitman-Agaude test – Hudson et al. (1987), McDonald-Kreitman test – McDonald & Kreitman (1991). One of the earliest and still wildly used method, the so called relative rate test (Miyata & Yasunaga 1980; Yang & Bielawski 2000), is frequently used to estimate the direction and strength of selection, especially to screen genes or regions of genes for positive selection, often in a genome wide manner (CSAC 2005; Bakewell et al. 2007; Kosiol et al. 2008). This test frequently relies on the assumption that synonymous sites evolve neutrally (Kimura 1968; King & Jukes 1969). But are synonymous sites really neutral?

1.1.4 Neutral standard

While it is unrealistic to assume that any of the mentioned sequence types would evolve strictly neutrally, intuitively, the category with the highest *d* is likely to contain the least number of functional sites. Recent comparative genomics studies on mammalian taxa indeed found, that when all sites are included in the analyses, synonymous sites evolve at the highest rate (Keightley et al. 2005; Gaffney & Keightley 2006). Another reason that supports the belief that synonymous sites evolve neutrally in mammals is namely their low N_e values, especially in primates (around 10^4 in primates, 10^5 in rodents, compared with 10^6 in fruit flies; see Mank et al. 2010), suggesting that most of the sites in mammals should either be neutral or just very weekly selected (Ohta 1995). Nevertheless recent evidence suggests that at least some mutations at synonymous sites, even in mammals, are selected against (Chamary et al. 2006; Drummond & Wilke 2008). As a consequence sequence types other than synonymous sites have been chosen for reference in tests of neutral evolution (e.g. Keightley et al. 2005; Gaffney & Keightley 2006b).

1.1.5 Context dependent mutational biases

Another important feature of mammalian genomes, which is known to have a strong impact on their evolution, is the mutational process, frequently referred to as context-dependent mutations. It was found by Josse et al. (1961) that the frequency of CpG dinucleotides in vertebrates is much lower than is expected from their base composition, most likely as a consequence of the so called CpG methylation deamination process (Bird 1980). Methylated CpG dinucleotides mutate to TpG or CpA (on the reverse strand) at a frequency which is 8-18 fold higher than the substitution rate observed at single nucleotides (Lunter & Hein 2004; Siepel & David Haussler 2004; Arndt & Hwa 2005; Duret & Arndt 2008). The effect of CpG hypermutability should inflate divergence rate estimates proportional to the site frequency, but this should also depends on the level of methylation. For example, flanking intergenic, and untranslated regions upstream of genes are known to contain CpG islands (Bird 1986; Takai & Jones 2002), with very high frequencies of CpG dinucleotides, which are largely unaffected by CpG hypermutability, most likely due to lower level of methylation in the upstream region of genes (Bird 1986). As a consequence of context dependency, when divergence is estimated at sites that are unlikely to be ancestrally part of CpG dinucleotides (Keightley & Gaffney 2003; Meunier & Duret 2004), rates are in general significantly reduced, and many sequence types present higher substitution rates than the rate observed at synonymous sites in mammals (Keightley et al. 2005; Gaffney & Keightley 2006), supporting the view that some of the mutations at synonymous sites are selected against (Chamary et al. 2006).

1.1.6 GC content evolution and biased gene conversion

There are at least two additional factors which are known to affect genome evolution. As large scale variation was found in the GC content of genomes between different taxonomic groups, Sueoka (1962) suggested that the underlying process of nucleotide sequence evolution is determined by substitution rate differences and derived the equations describing the dynamics of base compositional changes for

sequences which are not at compositional equilibrium. A possible consequence of the substitution rate differences is the so called isochore structure of genomes (Bernardi 2000), which are long genomic regions of relatively invariable GC content.

It was shown recently, that mammalian sequences are not at compositional equilibrium (Meunier & Duret 2004; Duret & Arndt 2008), and they evolve towards lower equilibrium GC contents (p_e). If this is the case then sequences of different GC contents (p_0) may evolve at different rates purely as a reason of their compositional differences. This may cause biases in estimates of strength of selection, which is frequently measured by comparing the rate of evolution of a sequence to the rate observed in an assumed neutral sequence, i.e. relative rate test.

Another process beside mutation and selection which affects the GC content of genomes is called biased gene conversion (BGC). This process is linked to DNA repair mechanisms which favour G or C nucleotides over A or T nucleotides in heteroduplexes, frequently occurring in crossovers during meiosis. BGC therefore may change the allele frequency and this process resembles to the action of selection favouring GC alleles (Nagylaki 1983).

1.2 Aims of this study

In this study, selective constraint is estimated on different sequence types of mammalian genomes, with the aim to infer the fraction of mutations which had been selectively eliminated since the speciation of a species or species pair. Selective constraint reflects, to some extent, the fraction of nucleotides which are under selection in a region or in a whole genome, and is an important parameter for estimating the genome-wide deleterious mutation rate (U) (e.g. Keightley & Eyre-Walker 2000) and the mutation load (L) at the population level (Kimura & Maruyama 1966). Following the estimates of U and L , the assumption of neutral evolution at synonymous sites is tested and consequences of the results are discussed.

Finally, the bias caused by non-equilibrium processes on inferences of selection is quantified in primates.

1.2.1 Genome wide selective constraint and the deleterious mutation rate

In the first results chapter, the rates of evolution at different sequence types (e.g. introns, intergenic regions, synonymous and non-synonymous sites) are compared, and by choosing the most likely candidate for a neutral standard, sequence specific selective constraints are estimated. Constraint is defined as the fraction of mutations which were selectively eliminated due to their fitness effects, and can be thought of as the lower limit on the fraction of functional sites in a given sequence type (i.e. when it is assumed that mutations are either strongly deleterious or completely neutral). By knowing the proportion of different site types in the genomes, I obtain an estimate of the genome wide constraint (C). With the estimate of μ from the neutral standard and C, it is possible to estimate the deleterious mutation rate per diploid genome per generation, which is known to be an important parameter in certain evolutionary models concerning the evolution of diploidy, sexual reproduction and recombination (e.g. Charlesworth & Charlesworth 1998; Keightley & Eyre-Walker 2000). The variation in constraint between different sequence types within and between species is also discussed, and I present some results that seem to contradict to the expectations postulated by the nearly neutral theory of molecular evolution (Ohta 1995).

1.2.2 Variations in synonymous and non-synonymous constraint in vertebrates

Second, as constraint exists at synonymous sites, and the two estimates for hominids and murids are different, therefore I explore the variation in constraint at synonymous and non-synonymous sites in eleven closely related vertebrate species pairs. Since the non-synonymous to synonymous rate ratio (d_N/d_S) is known to correlate with the divergence time (Rocha et al. 2006; Wolf et al. 2009), and N_e (Ellegren 2009), therefore they are likely to correlate with estimates of constraints. In this chapter, I test the correlation and present the effect of a covariate which affects

both N_e and the divergence time. The effect of synonymous constraint on the d_N/d_S-ratio is also discussed.

1.2.3 Effect of non-equilibrium processes on estimates of direction and strength of selection

It has been suggested that compositional differences among different sequence types, and the fact that mammalian genomes evolve towards a lower equilibrium (Piganeau et al. 2002; Duret & Arndt 2008), may lead to biases on the estimates of direction and strength of selection. In this chapter the effect of these so called non-equilibrium process are quantified, based on a deterministic model for GC content evolution (Sueoka 1962). As these processes may seriously influence estimates of selection based on the d_N/d_S ratio, especially in cases when selection is assumed to be relaxed (e.g. after gene duplication events which has been long assumed to play an important role in genome evolution an speciation; Lynch 2006; Hahn et al. 2007; Han et al. 2009) I test the bias caused by the non-equilibrium processes on the d_N/d_S ratio assuming an equilibrium value of 0.37 (Duret & Arndt 2008) and discuss the consequences.

2 Selective constraints and the deleterious mutation rate

The work described in this chapter is published (Eory, L, Halligan, D.L. and Keightley, P.D. 2010). LE designed the experiments, collected and analysed the data. PK and DH gave comments on various forms of the MS. This chapter has the same format as the published paper and the text herein contains only slight modifications.

2.1 Abstract

Protein-coding sequences make up only about 1% of the mammalian genome. Much of the remaining 99% has long been assumed to be junk DNA, with little or no functional significance. Here I show that in hominids, a group with historically low effective population sizes, all classes of non-coding DNA evolve more slowly than ancestral transposable elements, and so appear to be subject to significant evolutionary constraints. Under the nearly neutral theory, I expected to see lower levels of selective constraints on most sequence types in hominids than murids, a group that is thought to have a higher effective population size. I found that this is the case for many sequence types examined, the most extreme example being 5' UTRs, for which constraint in hominids is only about one-third that of murids. Surprisingly, however, I observed higher constraints for some sequence types in hominids, notably four-fold sites, where constraint is more than twice as high as in murids. This implies that more than about one-fifth of mutations at four-fold sites are effectively selected against in hominids. The higher constraint at four-fold sites in hominids suggests a more complex protein-coding gene structure than murids, and indicates that methods for detecting selection on protein coding sequences (e.g., using the d_N/d_S ratio), with four-fold sites as a neutral standard, may lead to biased estimates, particularly in hominids. My constraint estimates imply that 5.4% of nucleotide sites in the human genome are subject to effective negative selection, and that there are three times as many constrained sites within non-coding sequences as

within protein-coding sequences. Including coding and non-coding sites, I estimate that the genomic deleterious mutation rate, *U*, is equal to 4.2. The mutational load predicted under a multiplicative model is therefore about 99% in hominids.

2.2 Introduction

Among the most interesting questions to have arisen from the sequencing of complete genomes is the location and nature of functional sites in the genome. Protein-coding genes are one well-characterised class of functional sites, of which there are ~20,000 in mammals (Lynch 2007). However, protein-coding sequences make up only about 1% of the genome in mammals, and the extent of functional sites in non-protein-coding DNA is less well understood.

Functional sites can be recognised by their tendency to have lower levels of polymorphism and between species divergence than neutrally evolving segments of the genome (e.g., see Eyre-Walker & Keightley 1999; Andolfatto 2005; CSAC 2005). This results from the fact that most non-neutral new mutations are expected to disrupt function, and are therefore subject to purifying selection. Under the assumptions that beneficial mutations are rare and that mutations within a functional sequence are either neutral or strongly deleterious, one can estimate the fraction of selectively constrained sites from the difference in evolutionary divergence between the functional and an unconstrained sequence (Kondrashov & Crow 1993; Eyre-Walker & Keightley 1999).

The number of functional sites in the genome can then be used to estimate the genomic deleterious mutation rate (*U*). This is an important parameter in several evolutionary models, including the evolution of diploidy and the evolution of sex and recombination (Kondrashov 1988; Charlesworth & Charlesworth 1998). *U* can be estimated from the product of the neutral mutation rate per generation (μ), the fraction of selectively constrained sites in the genome (referred to as the genomic selective constraint, *C*) and the number of bases in the diploid genome (Kondrashov & Crow 1993). If the value of *U* for a species is much greater than one, it has been

argued that a species may be vulnerable to extinction as a consequence of genetic degradation brought about by the accumulation of new deleterious mutations (Muller 1950; Kimura & Maruyama 1966). Estimates of U of 0.91 in rodents (Gaffney & Keightley 2006) and 1.2 in fruitflies (Haag-Liautard et al. 2007) support the hypothesis that U is around one in some species. On the other hand, estimates in hominids have been 1.6 and 3.0 (Eyre-Walker & Keightley 1999; Nachman & Crowell 2000, respectively) for the protein-coding component of the genome, but are strongly affected by several assumptions. Foremost among these is the number of protein-coding genes, which has previously been greatly over-estimated. After rescaling using more recent estimates of 21,000 for the number of known protein-coding genes (Ensembl release 48), estimates for U for the protein-coding fraction of the genome, calculated as $U = 2 C_{0F} f_{0F} L N_g \mu$ (where f_{0F} and C_{0F} are the fraction of zero-fold degenerate sites and their level of constraint, respectively, L is the average length of protein coding sequences and N_g is the number of protein coding genes in the genome), are 0.4 and 0.7.

For several other reasons, however, these estimates of U ~0.5 are likely to be underestimates. First, they have omitted a net contribution from deleterious mutations in non-coding DNA. This is important because it has been estimated that rodent non-coding regions, for example, contain at least 4 times as many constrained sites as protein-coding sequences (IMGSC et al. 2002). The level of selective constraint operating in hominid non-coding sequences is subject to uncertainties. Constraint has been reported to be nearly absent in first introns of hominids (Keightley et al. 2005), whereas first introns appear to be among the most strongly constrained intronic regions in murids (Gaffney & Keightley 2006). Selective constraint has also been reported to be essentially absent in intergenic regions, when constraint was calculated using non-first introns as a neutral standard (Keightley et al. 2005). These findings are apparently at odds with the pilot phase of the Encyclopedia of DNA Elements project (ENCODE), which estimated that 5% of the human genome is under purifying selection and, that 60% of these constrained sites

overlap with experimentally identified functional regions (Birney et al. 2007). Given that less than 1% of the human genome codes for protein sequences (Ensembl release 48), this suggests that there are at least twice as many functional sites outside protein-coding sequences as within protein-coding sequences.

Second, estimates may depend on the chosen neutral standard. Previous analyses of non-degenerate sites in hominids have assumed that synonymous sites evolve neutrally (Eyre-Walker & Keightley 1999; Nachman & Crowell 2000); however, there is evidence that, even in mammals, four-fold degenerate sites are under some negative selection (Parmley et al. 2007; Chamary & Hurst 2005; Drummond & Wilke 2008).

Third, the level of selective constraint on non-synonymous sites has previously been estimated using small samples of as few as 50 protein-coding genes, which may be non-representative samples. Results vary substantially, from a value as low as 0.38 (Eyre-Walker & Keightley 1999) to 0.75 (Ohta 1995; Keightley et al. 2005), but may still be underestimates if the chosen neutral standard (non-first introns, or four-fold sites) is subject to selective constraint.

Fourth, it is known that mutational biases, especially CpG hypermutability, strongly affect evolutionary rate estimates for vertebrates (Arndt et al. 2003; Lunter & Hein 2004), and constraint estimates are biased if CpG hypermutability is unaccounted for (Gaffney & Keightley 2008).

Finally, constraint estimates may be biased if the model of sequence evolution is inadequate. For example, if the GC content of the neutral reference sequences are not at equilibrium, as is the case for transposable elements (IHGSC 2001), then the inferred mutation rate may be biased. Since transposable elements within the human genome have higher GC content than the estimated equilibrium GC content (IHGSC 2001) it is necessary to account for these differences when substitution rates are estimated. Conversely, if the equilibrium GC content is lower than the assumed neutral standard sequence, then the rate of change from GC→AT may be

underestimated. As a consequence, the expected number of changes and the estimated constraint level of a sequence would be over- or underestimated, depending on whether the sequence has a lower or higher GC content than the neutral standard. It is possible to correct for these differences if the equilibrium GC content is known (Halligan et al. 2004). Previous studies of transposable elements and other non-coding sequences found an excess of G or C to A or T (GC→AT) substitutions over A or T to G or C (AT→GC) substitutions (Arndt et al. 2003; Meunier & Duret 2004; Duret & Arndt 2008; but see also Webster et al. 2003). Based on these substitution rates, the equilibrium GC content of hominids has been estimated to be in the range 0.35 to 0.45 (Arndt et al. 2003; Meunier & Duret 2004; Duret & Arndt 2008).

In addition to the problems discussed above, another possible source of inaccuracy in estimating selective constraints comes from issues associated with alternatively spliced genes, which have been omitted from a previous analysis (Gaffney & Keightley 2006). Alternative splicing is believed to be prevalent in mammals, potentially operating in more than 60% of human genes (Johnson et al. 2003; Wang et al. 2008). Four-fold sites in alternatively spliced exons are under stronger purifying selection than those within constitutive exons (Ramensky et al. 2008), and proper splicing requires the presence of exonic splice enhancer (ESE) and silencer (ESS) sequences (Parmley et al. 2007), as well as other alternative splicing specific factors, frequently located in introns (Sorek 2003; Havlioglu et al. 2007). For these reasons, genes with alternatively spliced variants may be associated with a higher number of constrained sites than single transcript genes, and failure to account for multiple transcript genes may therefore lead to downwardly biased genome-wide constraint estimates.

Although mammalian four-fold sites have frequently been assumed to evolve neutrally (e.g., Kimura 1983; Eyre-Walker & Keightley 1999), here I use, in common with others (Chiaromonte et al. 2003; IMGSC et al. 2002), ancestral repeats (ARs), which are transposable elements (TEs) inserted into the common ancestor of

the species under consideration as a paradigm for neutrality. I do so for the following reasons: First, ARs are widespread in mammalian genomes; for example, they comprise 45% of the human genome (Ensembl release 48), and are sufficiently scattered as to serve as local neutral standards. Second, the distribution of insertion and deletion (indel) events in human and mouse ARs fits a neutral indel model, while the genome as a whole appears to be under indel purifying selection (Lunter et al. 2006). Third, my results show that evolutionary rates of ARs in hominids are very close to that observed in pseudogenes, which are usually assumed to evolve free of constraints.

Although here I assume that mutations within ARs are neutral, many mutations in mammalian non-coding regions are probably weakly selected. The fate of these mutations depends not only on random genetic drift, but on selection. The nearly neutral theory of molecular evolution (Ohta & Gillespie 1996) predicts that such mutations behave close to neutrally if their selective disadvantage (s) is less than $1/2N_e$, where Ne is the effective population size. Recently N_e has been estimated to be ~20,000 for hominids and ~600,000 for murids (Halligan et al. 2010), although the latter is likely to be an overestimate if synonymous sites, used to infer the neutral mutation rate, evolve under constraint. For hominids and murids, there is a range of selection coefficients $1/(2N_{e(murid)})<1/(2N_{e(hominid)})$ for which mutations are predicted to behave as effectively neutral in hominids but be selected against in murids. The nearly neutral theory therefore predicts lower levels of constraint in species with smaller N_e, and this trend has been observed for zero-fold degenerate sites and 5' and 3' flanking regions of hominid and murid genes (Keightley et al. 2005; Nikolaev et al. 2007). Here, I present some unexpected differences in constraint between these two taxonomic groups.

In this study, I aim to address the following questions: First, what fraction of the mammalian non-coding genome is under purifying selection? Second, how does the level of selective constraint vary between different sequence types? Third, are there differences in the level of selective constraint for specific sequence types between

murids and hominids? Fourth, are alternatively spliced genes associated with more or fewer constrained nucleotides? Finally, what is the level of genomic selective constraint and the genomic deleterious mutation rate in hominids?

2.3 Materials and methods

2.3.1 Data

Human and mouse genome sequence data (hg18, mm9) were downloaded from the University of California Santa Cruz (UCSC) bioinformatics website. A list of known human and mouse genes and transcripts and the corresponding annotations were obtained from a locally installed and modified mirror of the Ensembl MySQL core databases for human and mouse (Hubbard et al. 2007) (release 48 for human and release 49 for mouse, see also the entity relationship diagram Figure 2.1) and used for reference throughout the analysis. Putatively orthologous sequences, shared between human and chimpanzee or mouse and rat (i.e., genes that fulfilled my system of criteria for orthology described below), based on the human and mouse annotations and on BLASTZ (Schwartz et al. 2003) chained alignments (hg18 versus panTro2 and mm9 versus rn4 for hominids and murids, respectively), were obtained and analysed.

2.3.2 Mapping

A simple mapping procedure was utilised to allow the analyses of both single transcript (ST) and alternatively spliced (AS) genes. I stored values for three variables for each nucleotide in the analysed genomes for the following categories: 1. sequence category (coding, 5'/3' untranslated regions, intron, pseudogene, RNA coding gene), 2. coding information (zero-fold, four-fold degenerate sites) 3. repeat information (SINE, LINE, DNA, LTR or other repeat). Each of these three variables were allowed to take multiple values from their corresponding category, which were then used to track the annotations for ST and AS genes. I did not consider short tandem repeats, microsatellites, and RNA coding genes in the analysis, because the

sequencing and/or the alignment of these regions are problematic, or because they make up a relatively small portion of the genome.

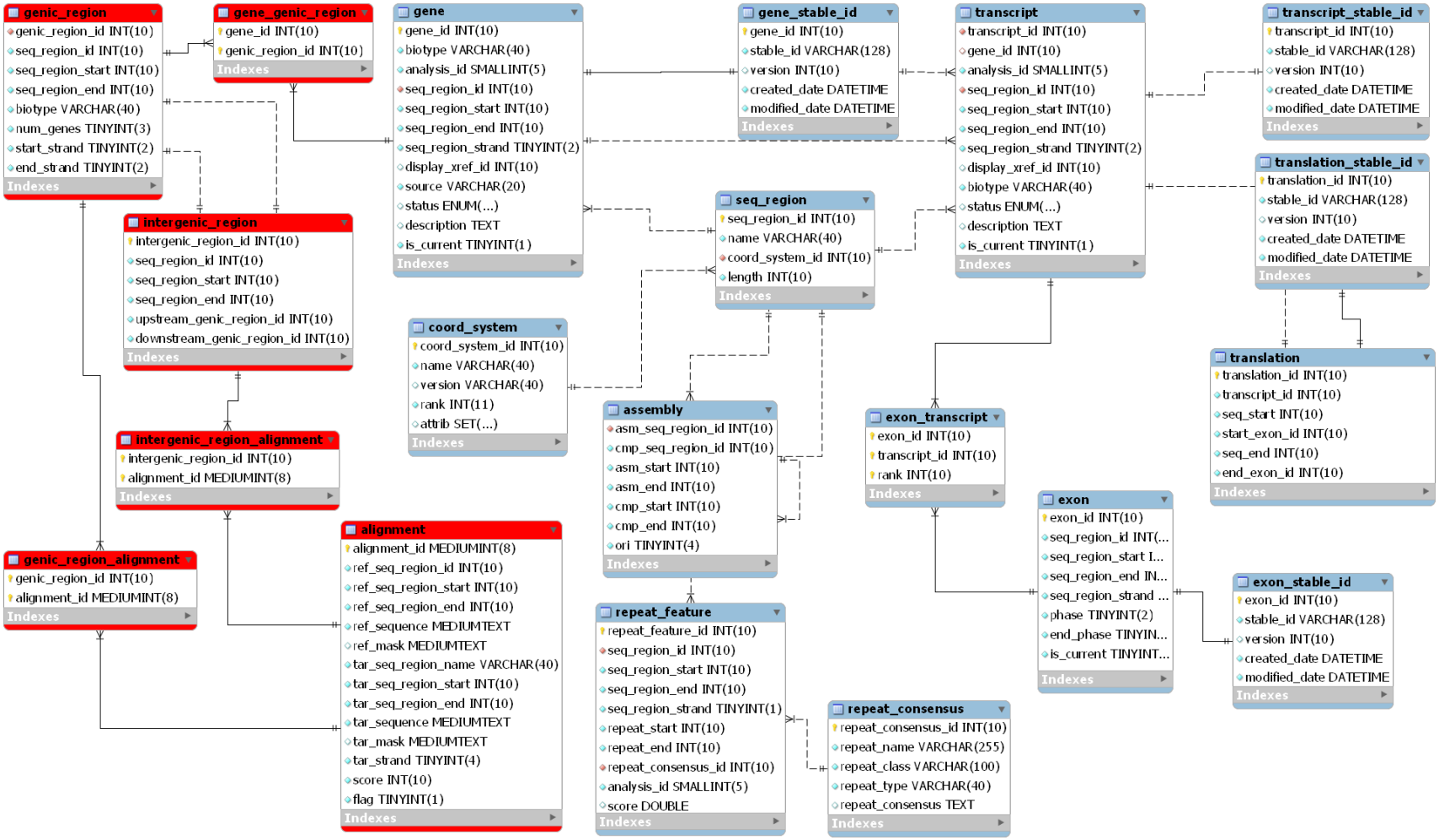

Figure 2.1. Entity relationship diagram of the modified Ensembl core database (represented with blue table heading). Tables with red heading were added to the databases and store the alignments and sequence quality information for genic and intergenic regions (see Material and Methods). Only tables relevant to the analysis are shown.

Intergenic regions (IG) and alternatively spliced sequence types considered in this analysis are the following: 1) 5' IG ≤ 5kb: proximal 5' intergenic region within 5kb of transcription start position; 2) 5'IG > 5kb: distal 5' intergenic region, more than 5kb from transcription start; 3) 3'IG ≤ 5kb: proximal 3' intergenic region within 5kb of transcription end; 4) 3'IG > 5kb: distal intergenic region more than 5kb from transcription end; 5) zero-fold – intron: non-degenerate sites alternatively spliced with intron; 6) zero-fold – 5'UTR: non-degenerate sites alternatively spliced with 5' untranslated regions (UTR); 7) zero-fold – 3'UTR: non-degenerate sites alternatively spliced with 3'UTRs; 8) 5'UTR – intron: 5'UTR alternatively spliced with intron; 9) 3'UTR – intron: 3'UTR alternatively spliced with intron. 5' and 3' intergenic regions were assigned to the corresponding categories by splitting the intergenic sequences into halves.

2.3.3 System of criteria for orthology

In the analysis, only transcripts that satisfied the following criteria were accepted as orthologous in the corresponding species (chimpanzee in the hominid and rat in the murid analysis). First, each exon of the transcript needed to be aligned to a homologous sequence in the corresponding BLASTZ alignment. Second, aligned exons were considered to be valid only if, for each exon, the homologous chromosomes, strands and coordinates indicated conserved synteny. Third, transcripts were excluded if they contained exons with frameshift mutations and/or premature stop codons. Fourth, transcripts that did not start/end in a start/stop codon in the two species were rejected. Genes were considered to be valid if they contained at least one valid transcript in the reference genome (human or mouse). Aligned intronic and flanking intergenic sequences that were not syntenic, where properties of synteny were defined using the valid aligned exons (i.e. aligned intergenic sequences were from the same chromosome and strand as the neighbouring exon, and were in proper order), were also left out of the analysis.

2.3.4 Sequence validity and alignment masking

While the human genome sequence is considered to be essentially complete (Human Build 36.1), the currently available draft assembly of the chimpanzee genome (Chimpanzee Build 2) may still contain an appreciable number of sequencing errors. To avoid a contribution from errors to my constraint estimates, I rejected those nucleotides from the analysis that had a quality score less than 40 (CSAC 2005). Similar to previous analyses, a masking protocol was also used to exclude those sites that were likely to be nonorthologous between human and chimpanzee (Keightley et al. 2005; Keightley & Gaffney 2003). Divergence was calculated in sliding windows of 40 alignment columns. Any regions covered by 50 or more contiguous windows, within which divergence was higher than 0.1 or the window contained less than 50% valid aligned bases, were masked out from the alignments. Similarly, I used alignment masking in the murid analysis with a cut-off divergence of 0.3, to allow comparisons of constraint on murid sequence types with a previous study (Gaffney & Keightley 2006).

2.3.5 Gene expression

Human and mouse gene expression data, based on high-density oligonucleotide microarray experiments (U133A and GNF1H for human, GNF1M for mouse), were obtained from the Novartis Gene Expression Atlas (Su et al. 2004) for a subset of genes (7609 and 9130 for human and mouse, respectively). A gene was considered to be expressed if its expression level was higher or equal to the dataset median (Vinogradov & Anatskaya 2007). Signals from probes representing the same tissue and the same genes were averaged, and three measures of expression (i.e. mean expression, maximum expression and expression breadth) were calculated for the genes over the 31 tissues common to both human and mouse (see Yang et al. 2005, but skin was also included). Genes were split into two subgroups with equal numbers (i.e. with low and high mean and maximum expression and with narrow and wide expression breadth), and divergence and abundance of intronic ARs were estimated

for these groups, where abundance is given as the ratio of total length of aligned TEs and the total length of introns.

2.3.6 Data analysis

The total length of each sequence category for ST and AS genes in the genome was based on the human annotation for hominids and on the mouse for murids. Substitution rates were estimated for each of the sequence types using the Kimura two-parameter model for multiple-hits correction (Kimura 1980). Evolutionary rates are given separately for all sites and for non-CpG-prone sites (sites not preceded by C or followed by G in either species); I used the latter method to avoid obtaining downwardly biased estimates of divergence at those sites that were not ancestrally part of a hypermutable CpG dinucleotide due to miscategorisation of mutations (Gaffney & Keightley 2008). Divergences at four-fold sites were calculated for codons where both aligned codons code for the same amino acid and at most a single change had occurred. I assume throughout this study that intronic and intergenic ARs evolve free of evolutionary constraints, so their evolutionary rates can be used to estimate the mutation rates for any sequence type. My analysis allows for the possibility that transcription associated processes may affect the mutation rate (Green et al. 2003; Majewski 2003), so different mutation rate estimates are used for the transcribed and the untranscribed portions of the genome, assuming that intergenic regions are not transcribed in the germline. The calculation of constraint was done using an extension of a previous method of Kondrashov & Crow (1993) and results are given for non-CpG-prone sites. I estimated the frequencies of four types of nucleotide changes in my neutral standards, by distinguishing between two pair-wise A↔T, C↔G, and two directional rates AT→GC and GC→AT, following the method of Halligan et al. (2004). The method was tested by simulations (Daniel L. Halligan, unpublished results). This method assumes that the directional rates depends on the equilibrium GC content, and splits the total number of observed AT↔GC ($N_{AT\leftrightarrow GC}$) changes into AT→GC ($N_{AT\rightarrow GC}$) and GC→AT ($N_{GC\rightarrow AT}$) changes as

follows: $N_{AT\to GC}=\frac{N_{AT\leftrightarrow GC}\,p_e(1-p_a)}{p_e(1-p_a)+(1-p_e)\,p_a}$ and $N_{GC\to AT}=\frac{N_{AT\leftrightarrow GC}(1-p_e)\,p_a}{p_e(1-p_a)+(1-p_e)\,p_a}$

where p_e is the equilibrium GC content, and p_a is the current GC content of the neutral standard, which is calculated after removing CpG-prone sites. By dividing the observed numbers of changes of the four types ($N_{AT\leftrightarrow TA}$, $N_{GC\leftrightarrow CG}$, $N_{AT\to GC}$, and $N_{GC\leftrightarrow AT}$) by the number of sites at which a change of given type could occur in one step, I estimate four corresponding substitution rates (k_i). These rates are then used to estimate the expected number of substitutions in an adjacent putatively functional sequence by the equation where m_i is the corresponding number of sites in the sequence of interest. Then by counting the number of substitutions (O) in the sequence type (seq) I calculate selective constraint as C_{seq}=1-O_{seq}/E_{seq} (Halligan et al. 2004). As the rate estimates are uncorrected for multiple hits the constraint estimates may be biased, the level of which depends on 1) the overall divergence of the species pair (e.g. estimates for human-chimp are less affected than for the more highly diverged mouse-rat pair) 2) the sequence type analysed (i.e. for sequences which evolve close to neutrally the effect is probably smaller than for sequences under higher level of constraint).

To investigate how constraint changes from the transcriptional start and end of genes into deep intergenic regions, I calculated average constraint in 400 nucleotide non-overlapping sliding windows. 95% confidence intervals for divergence and constraint were obtained by splitting each chromosome into 1 Mb blocks and bootstrapping 1,000 times by block (Keightley & Gaffney 2003).

2.3.7 Genomic deleterious mutation rate

Evolutionary rate estimates for intronic and intergenic ARs were used as estimates of the mutation rates for genic and intergenic regions, respectively. The genomic deleterious mutation rate per diploid per generation was estimated from the product of the genomic selective constraint per nucleotide and the number of point mutations in the repeat-free portion of the diploid genome. Genomic selective constraint (C) for the euchromatic genome was calculated by summing the products of constraint and

genomic sequence length for each sequence type for single transcript and alternatively spliced genes separately (Gaffney & Keightley 2006). In my study, the number of mutations (M) was the sum of the product of the repeat free genic length (outside CpG sites), the genic neutral mutation rate, and the repeat-free intergenic length (outside CpG sites) and intergenic neutral mutation rate, for regions associated with my ST and AS genes. M was corrected by taking into account mutations from CpG sites by considering CpG hypermutability, based on estimates already published (Arndt et al. 2003; Lunter & Hein 2004). The genomic deleterious mutation rate U was calculated as $U = CM$.

2.4 Results

In total, there are 21,108 known protein-coding genes in the human genome (Ensembl release 48), comprising 8,491 single transcript (ST) and 12,617 alternatively spliced (AS) genes, of which 15,696 fulfilled my criteria for orthology (see materials and methods). The mouse genome (Ensembl release 49) currently has 11,749 ST and 10,111 AS genes annotated, and in this study I analysed 14,410 orthologous genes.

2.4.1 Genome composition

In both taxa, mean coding length is around 1,500 nt per gene, although ST coding sequences tend to be shorter on average than AS sequences (mean = 1,200 and 1,800 nt, respectively). The primary cause of this ~50% difference is the presence of non constitutively spliced exons (i.e. coding regions having overlapping annotations with introns or with 5'/3' UTRs) (see Table 2.1A and 2.1B). 5' and 3' UTR sequences are also 50-70% longer in AS genes. Hominid AS genes contain 2.2 times more intronic sequence than ST genes, while in murids the difference is 2.5 fold. Although the composition of human and mice genes are similar in many respects, human genes tend to contain more transposable elements within introns (~30% more frequent in humans) and have longer UTRs (by around 20%).

Table 2.1. Sequence composition of coding regions in the genome.

A. Hominids

Sequence type	Single transcript		Alternatively spliced	
	***N* (Mb)**	***N*/gene**	***N* (Mb)**	***N*/gene**
zero-fold	8.1	767.6	9.3	889.2
four-fold	2.0	190.2	2.2	213.7
5'UTR	1.9	180.2	1.9	182.3
3'UTR	8.6	813.5	10.6	1,004.9
intron (non rpt)	180.1	16,989.8	399.6	38,043.6
intronic repeats	144.4	13,622.8	312.8	29,772.7
coding – intron	-	-	2.5	233.9
coding – 5'UTR	-	-	1.4	134.3
coding – 3'UTR	-	-	1.0	98.2
5'UTR – intron	-	-	1.0	97.3
3'UTR – intron	-	-	2.1	204.7
minor seq types	0.0	1.7	0.5	48.7
Total	345.3	32,565.7	745.1	70,923.5

B. Murids

Sequence type	Single transcript		Alternatively spliced	
	***N* (Mb)**	***N*/gene**	***N* (Mb)**	***N*/gene**
zero-fold	11.1	863.9	8.0	949.9
four-fold	2.7	211.9	1.9	228.9
UTR5	2.0	152.7	1.4	160.8
UTR3	9.2	719.2	7.9	940.4
intron (non rpt)	229.7	17,953.6	345.7	40,921.0
intronic repeats	117.9	9,212.5	167.0	19,773.9
coding – intron	-	-	1.6	190.7
coding – 5'UTR	-	-	0.8	90.6
coding – 3'UTR	-	-	0.7	88.7
5'UTR – intron	-	-	0.6	68.5
3'UTR – intron	-	-	1.5	172.4
minor seq types	0.0	1.3	0.3	30.6
Total	372.5	29,115.2	537.4	63,616.5

Only those sequence types whose length exceeds 1Mb are detailed. Minor categories (e.g., intronic pseudogenes/RNA coding genes, and other alternatively spliced categories) are shown under the “minor seq types” category. *N* (Mb) refers to the number of sites of the specified sequence type. *N*/gene is the mean number of sites in the category per gene. Overlapping annotations for sequence types associated only with alternatively spliced genes are shown as double categories.

2.4.2 Variation in evolutionary rates across sequences utilised as neutral standards in previous analyses

The calculation of constraint depends critically on the choice of neutral standard, so I compared mean nucleotide divergence among different sequence types that have been used in previous studies as a paradigm for neutrality. Mean divergences for various sequence types in hominids are shown in Figure 2.2. However, correlations between divergence and local GC content are known to exist in hominids (Duret & Arndt 2008), and could affect these divergences.

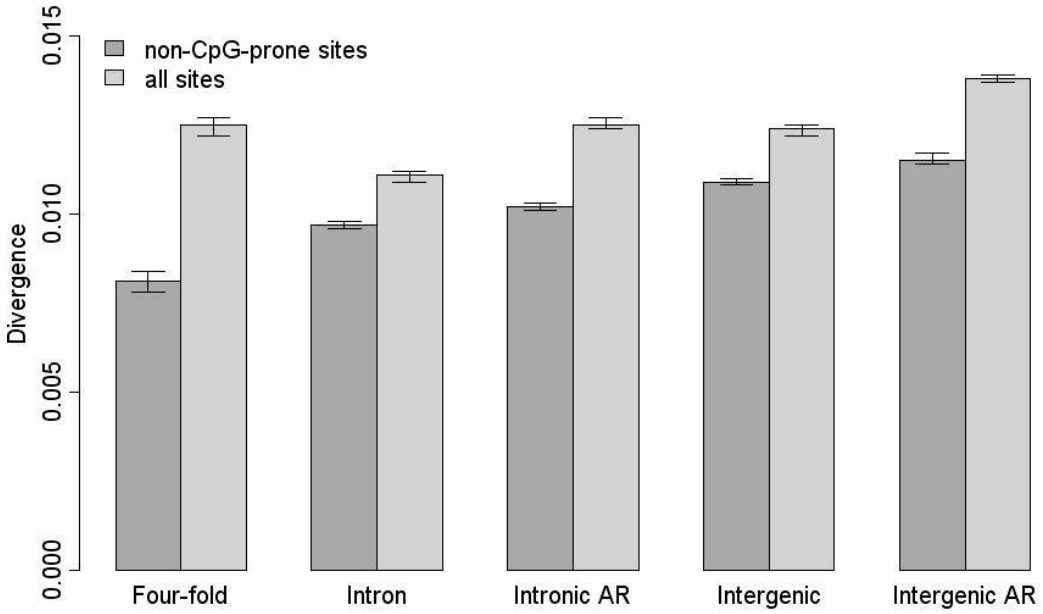

Figure 2.2. Mean nucleotide divergence in different sequence types in hominids that appear to evolve close to neutrally. Dark and light bars indicate divergence at non-CpG-prone and at all sites, respectively. 95% confidence intervals are indicated by bars.

To investigate this I plotted divergence against GC content for the different sequence types (Figure 2.3 and 2.4). The correlations between divergence and GC content appear to be non-linear, as illustrated by the locally weighted regression lines

(Lowess curves; Cleveland & Devlin 1988) of divergence on GC content, which are indicated in the figures. If all sites are included (i.e. both CpG-prone and non-CpG-prone are included) the highest rate of evolution is found in intergenic ARs and the lowest rate is found in introns. Mean divergences for four-fold sites, intronic ARs and unique intergenic regions are all very similar to each other, although four-fold site divergence is more strongly affected by the local GC content, and in regions where the GC content drops below 40% four-fold sites become the slowest evolving category (Figure 2.4).

Divergence for hominids in intergenic ARs (0.0138 [95% confidence interval = 0.0137, 0.0139],) is very close to an estimate from my data for pseudogenes (0.0141 [0.0135, 0.0148]), which are frequently assumed to evolve in a neutral manner (see e.g. Li et al. 1981). However, this latter estimate has broad confidence limits, due to the small number of annotated pseudogenes in Ensembl, and the ambiguity of the annotations (less than 4,500 annotated in total and less than 50 annotated as known).

Removing the effect of CpG hypermutability on divergence by considering non-CpG-prone sites only (see Materials and Methods) causes a marked reduction in all estimates (Figure 2.2) and there is a weaker correlation between divergence and GC content (Figure 2.3). The reduction in divergence is most pronounced at four-fold sites (35%). Estimates are lower by 20% for intronic ARs, 16% for intergenic ARs and 12% for non-repeat introns and intergenic regions.

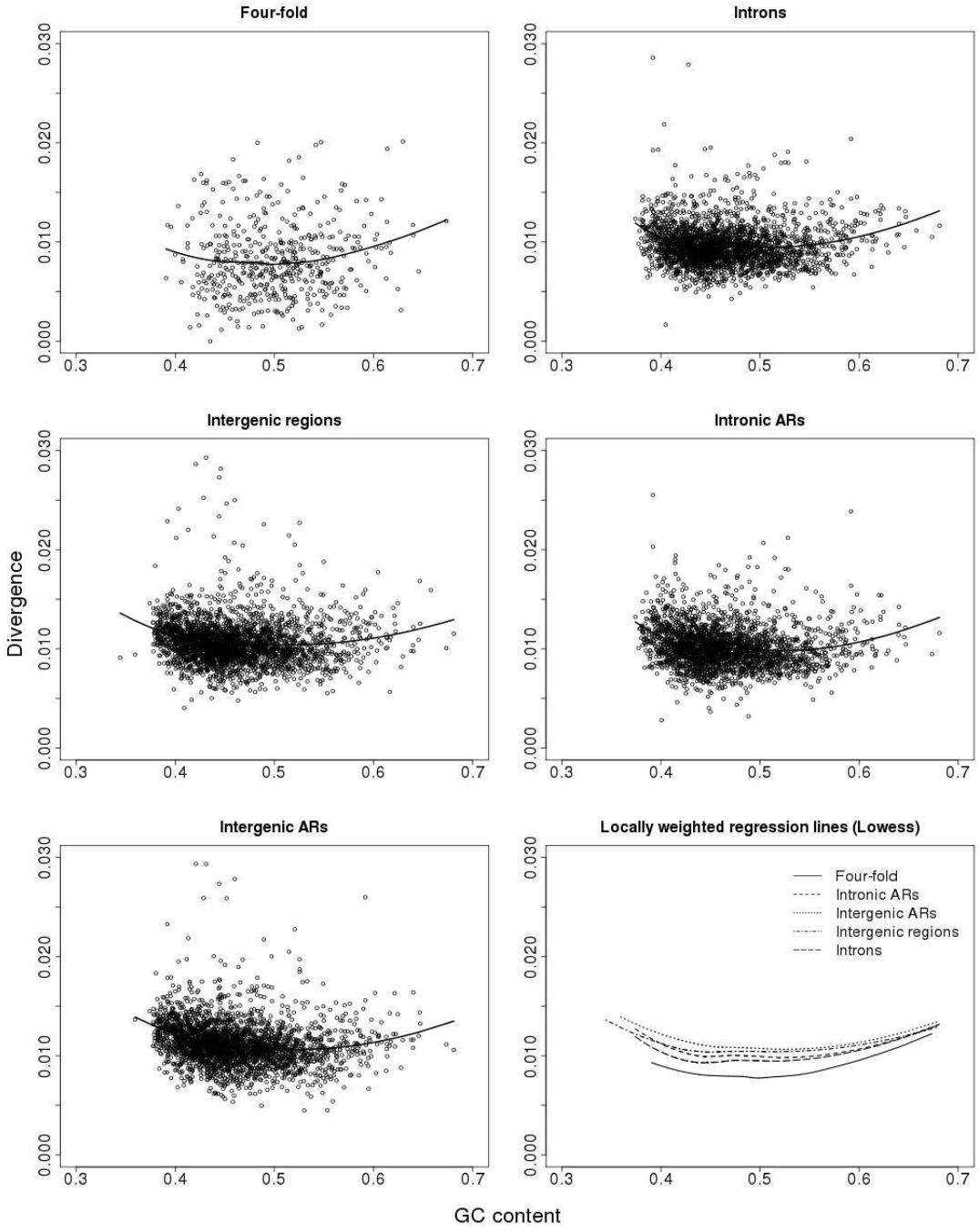

Figure 2.3. Correspondence between the current GC content of hominids and divergence estimates at non-CpG-prone sites. Divergence and GC content were calculated in 1MB segments along the chromosomes. Lowess curves (locally weighted regression lines) are compared in the bottom right panel. Removal of segments of <200 base-pairs causes a reduction of data points at four-fold sites.

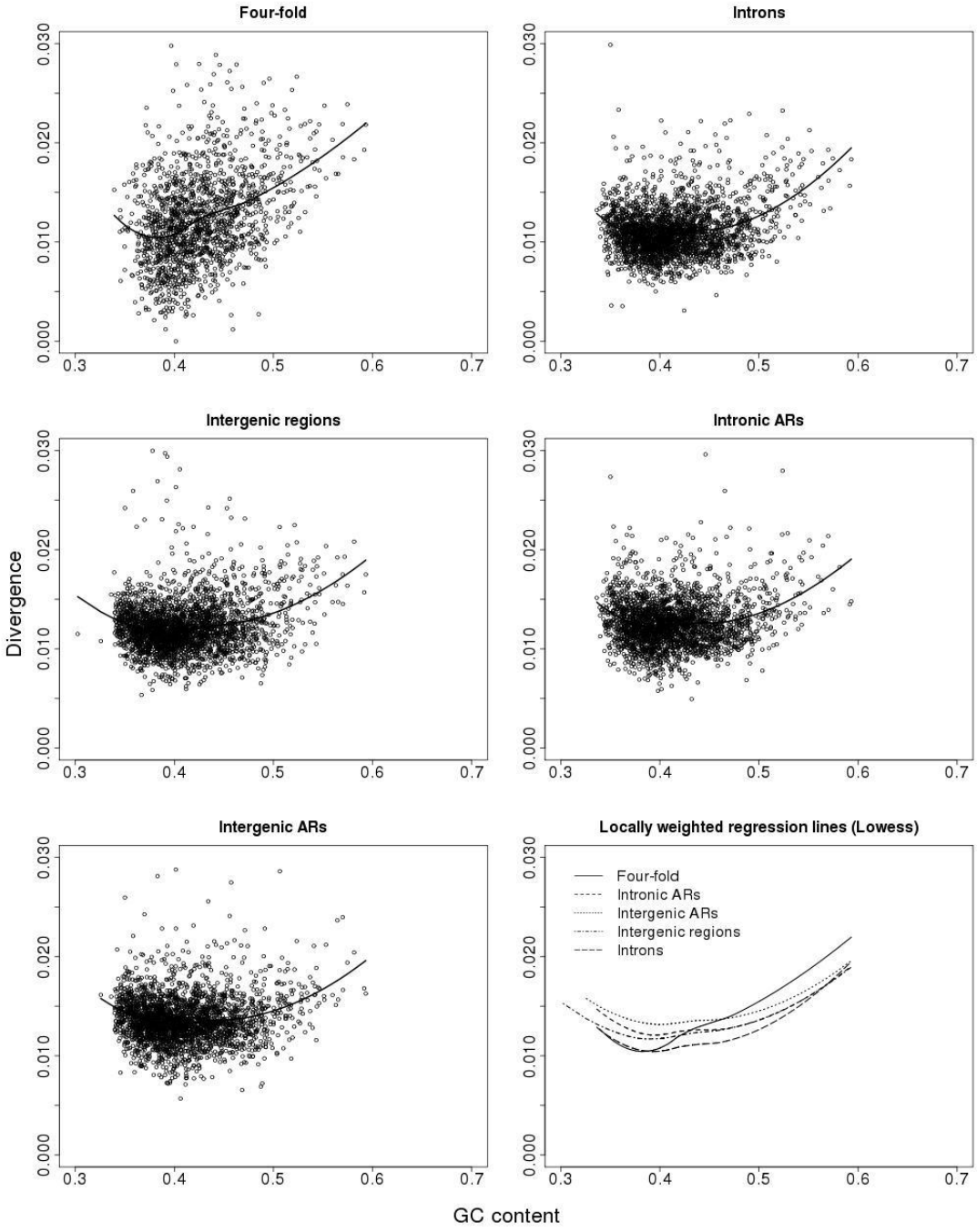

Figure 2.4. Correspondence between the current GC content of hominids and divergence estimates at all sites. Divergence and GC content were calculated in 1MB segments along the chromosomes. Lowess curves (locally weighted regression lines) are compared in the bottom right panel. Removal of segments of <200 base-pairs causes a reduction of data points at four-fold sites.

The pattern of differences in evolutionary rates at all and at non-CpG-prone sites reflects the different CpG contents of the sequences examined (Table 2.2), in particular the fact that coding sequences are enriched for CpGs. When comparing the rates of evolution at non-CpG-prone sites between sequence types, four-fold sites are by far the slowest evolving category. This conclusion is not affected by the regional GC content (Figure 2.3). The rate for four-fold sites is well below that of both introns and intergenic sequences. Intergenic ARs are the swiftest evolving category (0.0115 [0.0114, 0.0117]), and their rate is again similar to that observed in pseudogenes (0.0113 [0.0108, 0.0119]). Intergenic and intronic ARs evolve at a higher rate than the corresponding unique intergenic and intronic sequences and this trend is unaffected by the underlying GC content, which may suggest that these sites evolve closer to neutrally (Figure 2.3).

Table 2.2. Percentage GC and CpG in different sequence classes in hominids.

Sequence type	All sites	Non-CpG prone sites	CpG (%)
	GC (%)	GC (%)	
zero-fold	48.8	54.81	5.81
four-fold	54.0	67.97	9.56
intron	39.0	43.36	1.82
intronic AR	42.9	48.33	2.35
intergenic	39.0	43.43	1.75
intergenic AR	41.4	46.6	1.86

GC content is shown for all genes, including AS and ST genes, at all and at non-CpG-prone sites, CpG composition is obtained by considering the human annotation only.

As noted previously in murids (Gaffney & Keightley 2006), ARs within introns exhibit slightly lower divergence than ARs located within intergenic regions. This may relate either to transcription specific biases, e.g. transcription coupled mutation

and repair (Green et al. 2003; Majewski 2003), or to a higher frequency of functional sites (e.g., regions responsible for maintaining stable mRNA structure or binding sites for regulatory proteins or small RNAs).

Sequence divergence, at least in protein-coding sequences, is known to be correlated with gene expression level in vertebrates (Subramanian & Kumar 2004; Drummond & Wilke 2008), while there are results suggesting that intron length is selected against in highly expressed genes (Castillo-Davis et al. 2002; Urrutia 2003). If the divergence of intronic ARs differs between genes of high and low expression, and there is a difference in the fraction of ARs within introns between highly and lowly expressed genes, then estimates of neutral rates, summed across genes, may be biased and as a consequence may affect the estimated constraint. To test for the magnitude of difference in AR divergence and abundance, I analysed a set of genes in hominids and murids for which expression data are available (Su et al. 2004). Consistent with results for protein coding sequences, I find that intronic ARs evolve at a slightly higher rate in genes with lower expression level than in genes with higher level of expression (Table 2.3) in both species. Furthermore, consistent with the theory of selection for shorter introns in highly expressed genes, AR abundance is also lower in genes with higher expression levels. The observed differences in divergence between estimates for highly and lowly expressed genes could either be caused by mutational biases or by selection but these differences are small and non-significant. This suggest, that any error in constraint estimates caused by differential expression is likely to be small.

Table 2.3. Intronic AR abundance and evolutionary rate in genes with high or low expression patterns in human and mouse.

A. Human

Measure of Expression		Expression level		P
		High/broad	Low/narrow	
maximum	AR abundance	0.3726	0.3987	0.20
	divergence	0.0098	0.0101	0.54
mean	AR abundance	0.3775	0.3956	0.43
	divergence	0.0097	0.0102	0.35
breadth	AR abundance	0.3840	0.3912	0.76
	divergence	0.0097	0.0102	0.56

B. Mouse

Measure of Expression		Expression level		P
		High/broad	Low/narrow	
maximum	AR abundance	0.1951	0.2054	0.37
	divergence	0.1528	0.1562	0.67
mean	AR abundance	0.1982	0.2026	0.71
	divergence	0.1521	0.1568	0.53
breadth	AR abundance	0.2018	0.1993	0.82
	divergence	0.1521	0.1570	0.59

AR divergence is given at non-CpG-prone sites. P-values for differences in constraint between ST and AS genes were estimated based on the bootstrap replicates assuming mean difference of 0.

2.4.3 Variation in selective constraints between coding and non-coding sequences

In order to assess constraint variation among different sequence types and to test for possible differences in constraint levels between ST and AS genes, I estimated constraints for various sequence categories using intronic and intergenic ARs as my neutral standards for genic and intergenic sequence evolution, respectively. The higher GC content of ARs relative to the observed GC content of introns and IG

regions (see Table 2.2) suggests that ARs are not at compositional equilibrium and, as a consequence, the estimated neutral substitution rates might be overestimated, which in turn may bias constraint estimates. This is further supported with my observation that the relationship between divergence and GC content is non-linear and that the locally weighted polynomial regression lines suggest that sequences with extreme GC contents evolve the fastest (Figure 2.3 and 2.4).

I attempted to account for the deviation in GC content from the assumed equilibrium by using the method of Halligan et al. (2004) (see Materials and Methods), and assumed an equilibrium GC content of 0.37 in hominids (Arndt et al. 2003; Meunier & Duret 2004; Duret & Arndt 2008). Constraint was calculated for each of nine sequence types common to both ST and AS genes (zero-fold/four-fold degenerate sites, 5'/3' UTRs, introns, and proximal and distal 5'/3' flanking regions), along with an additional five sequence categories specific to AS genes (i.e. non-degenerate sites alternatively spliced with 5'/3' UTRs or introns, and 5'/3' UTRs alternatively spliced with introns). Estimates of selective constraint for the human-chimpanzee comparison are given for non-CpG-prone sites in Table 2.4.

My results suggest that a substantial fraction of each sequence type evolves under purifying selection. As expected, the highest constraint was observed at zero-fold degenerate sites, with values of 0.70 and 0.76 for ST and AS genes, respectively. Alternatively spliced zero-fold sites (i.e. when protein-coding annotation overlaps with 5'/3'UTRs or with introns) are also strongly constrained, with constraint in the range of 0.60 and 0.73. My estimates for four-fold degenerate sites suggest that a substantial number of sites within these regions evolve under purifying selection in hominids (constraint estimates are 0.22 and 0.27 on ST and AS genes, respectively). Constraint on 5'UTRs (ST: 0.15, AS: 0.19), is lower than on four-fold sites and on 3'UTRs (ST: 0.16 and AS: 0.22). Although low, selective constraint is evident in introns (ST: 0.03 and AS: 0.04). Constraint is also evident in intergenic regions, where it is higher in regions close to the transcription start and end positions of genes (around 0.09 for intergenic regions within 5 kb from transcription start and end) and

lower, but still significant, deeper into intergenic regions (around 0.04). Estimates of constraint are generally higher in AS genes than ST genes, significantly so for zero-fold and four-fold degenerate sites and 3'UTRs. The greatest difference is observed for 3'UTRs where constraint is 35% higher in AS genes, while for zero-fold and four-fold sites the differences are 9% and 27%, respectively.

Under the assumption of an equilibrium GC content of 0.40 (Khelifi et al. 2006), constraint estimates for the mouse-rat comparison differ somewhat from previous estimates for murid single transcript genes (Gaffney & Keightley 2006) (Table 2.5). This is at least partly caused by the fact that the method for calculating constraint accounts for the effect of a compositional difference from the equilibrium GC content. My results show that differences in constraints between AS and ST are more pronounced in hominids than murids. Differences are generally highly significant for genic sequence types, the only exception being at four-fold sites. I also found significant differences in constraint between intergenic categories in murids.

The differences in constraint between hominids and murids for each sequence type are highly significant in most cases ($P<0.01$). The only exceptions are proximal 5' intergenic regions and distal 5' intergenic regions in AS and ST genes, respectively. In general, constraints on murid sequence types are higher than for the corresponding hominid sequence types. The biggest difference is observed at 5'UTRs, where mean constraint reaches 0.48 in murids. Strikingly, I also found sequence types for which constraint in hominids exceeds that for murids. This is most pronounced at four-fold degenerate sites (difference = 0.11 and 0.16 for ST and AS genes, respectively) and proximal 3' intergenic regions (difference = ~0.06), although I also estimated slightly higher constraint levels for introns and proximal 5' intergenic regions.

Table 2.4. Mean constraint and mean number of constrained sites in different sequence types of Hominids.

	Single transcript			Alternatively spliced			P
Sequence type	**Constraint (95%CI)**	**N_c(Mb)**	**N_c/gene**	**Constraint (95%CI)**	**N_c(Mb)**	**N_c/gene**	
zero-fold	0.698 (0.685, 0.710)	5.68	536	0.759 (0.749, 0.768)	7.09	675	<0.001
four-fold	0.215 (0.185, 0.247)	0.43	41	0.272 (0.248, 0.295)	0.61	58	0.002
5'UTR	0.149 (0.117, 0.178)	0.28	27	0.187 (0.161, 0.213)	0.36	34	0.068
3'UTR	0.161 (0.146, 0.177)	1.39	131	0.216 (0.205, 0.228)	2.28	217	<0.001
intron (non rpt)	0.033 (0.028, 0.038)	5.96	562	0.039 (0.035, 0.042)	15.46	1472	0.126
zero-fold – intron	-	0	0	0.604 (0.576, 0.628)	1.48	141	
zero-fold – 5'UTR	-	0	0	0.733 (0.710, 0.755)	1.80	171	
zero-fold – 3'UTR	-	0	0	0.710 (0.685, 0.736)	1.00	95	
5'UTR – intron	-	0	0	0.164 (0.127, 0.199)	0.17	16	
3'UTR – intron	-	0	0	0.150 (0.122, 0.174)	0.15	15	
5' IG ≤ 5k	0.090 (0.079, 0.101)	1.99	188	0.099 (0.088, 0.110)	0.21	20	0.180
5' IG > 5k	0.038 (0.032, 0.044)	7.75	731	0.042 (0.036, 0.047)	7.75	738	0.184
3' IG ≤ 5k	0.088 (0.077, 0.099)	1.92	181	0.091 (0.080, 0.102)	1.66	158	0.778
3' IG > 5k	0.041 (0.033, 0.048)	7.90	745	0.036 (0.030, 0.042)	5.43	517	0.290

Data are presented for single transcript and alternatively spliced genes separately. Mean constraint was calculated for non-CpG-prone sites. N_c, N_c/gene refer to the number of constrained sites per category and the number of constrained sites per gene respectively.

Sequence category abbreviations for proximal/distal 5'/3' flanking regions are 5'/3' IG ≤/> 5kb. P-values for differences in constraint between ST and AS genes were estimated based on the bootstrap replicates assuming mean difference of 0.

Table 2.5. Mean constraint and mean number of constrained sites in different sequence types of Murids.

Sequence type	Single transcript			Alternatively spliced			P
	Constraint (95%CI)	N_C(Mb)	N_C/gene	Constraint (95%CI)	N_C(Mb)	N_C/gene	
zero-fold	0.795 (0.789, 0.802)	8.79	648	0.833 (0.827, 0.839)	6.68	806	<0.001
four-fold	0.108 (0.098, 0.117)	0.29	22	0.117 (0.107, 0.127)	0.23	27	0.168
5'UTR	0.465 (0.454, 0.477)	0.91	67	0.502 (0.489, 0.516)	0.68	82	<0.001
3'UTR	0.241 (0.232, 0.250)	2.22	163	0.294 (0.284, 0.304)	2.34	282	<0.001
intron (non rpt)	0.014 (0.010, 0.018)	3.26	240	0.029 (0.026, 0.033)	10.1	1217	<0.001
zero-fold – intron	-	-	-	0.702 (0.683, 0.719)	1.13	136	
zero-fold – 5'UTR	-	-	-	0.830 (0.811, 0.847)	0.64	77	
zero-fold – 3'UTR	-	-	-	0.804 (0.785, 0.824)	0.6	73	
5'UTR – intron	-	-	-	0.386 (0.362, 0.409)	0.22	27	
3'UTR – intron	-	-	-	0.173 (0.152, 0.194)	0.25	30	
5' IG ≤ 5k	0.071 (0.066, 0.076)	2.47	182	0.088 (0.082, 0.094)	1.5	181	<0.001
5' IG > 5k	0.042 (0.038, 0.045)	10.18	750	0.053 (0.048, 0.057)	7.55	910	0.002
3' IG ≤ 5k	0.023 (0.017, 0.028)	0.73	54	0.029 (0.022, 0.036)	0.45	54	0.168
3' IG > 5k	0.050 (0.046, 0.054)	12.16	896	0.061 (0.056, 0.066)	6.86	827	0.002

Data are presented for single transcript and alternatively spliced genes separately. Mean constraint was calculated for non-CpG-prone sites. N_C, N_C/gene refer to the number of constrained sites per category and the number of constrained sites per gene respectively.

Sequence category abbreviations for proximal/distal 5'/3' flanking regions are 5'/3' IG ≤/> 5kb. P-values for differences in constraint between ST and AS genes were estimated based on the bootstrap replicates assuming mean difference of 0.

In this study, I estimated constraint based on neutral substitution rates corrected for non-equilibrium processes. While genome-wide estimates of the equilibrium GC content for hominids are consistently around 0.37 (Arndt et al. 2003; Meunier & Duret 2004; Duret & Arndt 2008), the only estimate for murids, based on processed pseudogenes, is 0.40 (Khelifi et al. 2006). To assess the effect of deviation from the equilibrium GC content on my constraint estimates, I calculated constraint by changing the assumed equilibrium GC content from 0.25 to 0.50 in steps of 0.05 (Table 2.6). Constraint estimates depend on (a) the difference between the current and equilibrium GC content of the neutral standard and (b) the actual GC content of the sequence types examined. For sequences with a higher actual GC content than that of ARs, constraint estimates decrease as a function of increasing equilibrium GC content (e.g., four-fold sites and 5'UTRs). For sequences with lower GC content than ARs, the trend is reversed (e.g., 3'UTRs, introns and intergenic regions).

Here, I have assumed that the equilibrium GC content is constant throughout the genome. However, equilibrium GC content has been shown to be correlated with recombination rate and with the current GC content of the sequence (Meunier & Duret 2004; Duret & Arndt 2008). This might lead my estimates to be biased, and so could compromise my constraint comparison if most human genes, for example, are preferentially located within regions having a higher equilibrium GC content. It is, for example, known that broadly expressed genes are preferentially located in GC rich regions (Lercher et al. 2003). To test for the effects of recombination rate and local GC content on my estimates, I calculated constraint for hominids by using a multiple regression of the equilibrium GC content on local GC content and recombination rate (using information provided by Laurent Duret) to predict the local equilibrium GC content along the human genome. Crossover rates were taken from the HAPMAP genetic map (IHMC 2005) and were averaged over 1Mb segments. The resulting estimates based on the local equilibrium GC content are almost the same as obtained by assuming a constant stationary GC content of 0.37 (see Table 2.7).

Table 2.6. Constraint estimates on different sequence types of hominids and murids assuming different equilibrium GC contents.

A. Hominid

Sequence type	Equilibrium GC content (%)						
	25	**30**	**35**	**37**	**40**	**45**	**50**
zero-fold	0.737 (0.726, 0.747)	0.735 (0.724, 0.745)	0.732 (0.721, 0.743)	0.732 (0.720, 0.742)	0.730 (0.719, 0.741)	0.728 (0.717, 0.739)	0.726 (0.715, 0.737)
four-fold	0.289 (0.264, 0.313)	0.272 (0.245, 0.296)	0.253 (0.227, 0.279)	0.246 (0.219, 0.272)	0.234 (0.206, 0.260)	0.213 (0.186, 0.241)	0.191 (0.161, 0.219)
5'UTR	0.208 (0.180, 0.233)	0.192 (0.165, 0.218)	0.175 (0.148, 0.202)	0.168 (0.140, 0.196)	0.158 (0.131, 0.186)	0.140 (0.112, 0.168)	0.122 (0.093, 0.152)
3'UTR	0.179 (0.165, 0.192)	0.184 (0.171, 0.198)	0.190 (0.176, 0.204)	0.192 (0.179, 0.206)	0.195 (0.182, 0.209)	0.201 (0.187, 0.214)	0.206 (0.193, 0.219)
intron (non rpt)	0.017 (0.013, 0.021)	0.025 (0.021, 0.029)	0.034 (0.029, 0.038)	0.037 (0.033, 0.041)	0.042 (0.038, 0.046)	0.050 (0.046, 0.054)	0.057 (0.053, 0.061)
5' IG ≤ 5kb	0.094 (0.082, 0.106)	0.094 (0.083, 0.106)	0.094 (0.083, 0.106)	0.095 (0.083, 0.105)	0.095 (0.083, 0.106)	0.095 (0.084, 0.107)	0.095 (0.084, 0.107)
5' IG > 5kb	0.024 (0.019, 0.030)	0.031 (0.025, 0.036)	0.037 (0.032, 0.043)	0.040 (0.034, 0.046)	0.044 (0.038, 0.049)	0.050 (0.044, 0.055)	0.056 (0.050, 0.061)
3' IG ≤ 5kb	0.081 (0.070, 0.092)	0.085 (0.073, 0.096)	0.088 (0.078, 0.099)	0.089 (0.079, 0.100)	0.091 (0.079, 0.102)	0.094 (0.084, 0.105)	0.097 (0.086, 0.108)
3' IG > 5kb	0.025 (0.019, 0.032)	0.031 (0.024, 0.037)	0.036 (0.030, 0.043)	0.038 (0.032, 0.045)	0.042 (0.035, 0.048)	0.047 (0.040, 0.053)	0.052 (0.045, 0.058)

B. Murid

Sequence type	Equilibrium GC content (%)						
	25	**30**	**35**	**37**	**40**	**45**	**50**
zero-fold	0.814 (0.807, 0.820)	0.813 (0.807, 0.820)	0.813 (0.806, 0.819)	0.812 (0.806, 0.819)	0.812 (0.806, 0.818)	0.811 (0.805, 0.817)	0.810 (0.804, 0.817)
four-fold	0.175 (0.165, 0.184)	0.155 (0.146, 0.164)	0.134 (0.125, 0.143)	0.126 (0.116, 0.135)	0.112 (0.102, 0.121)	0.088 (0.078, 0.097)	0.062 (0.053, 0.072)
5'UTR	0.511 (0.498, 0.523)	0.502 (0.490, 0.514)	0.492 (0.479, 0.504)	0.488 (0.476, 0.500)	0.482 (0.470, 0.494)	0.471 (0.459, 0.484)	0.460 (0.448, 0.473)
3'UTR	0.254 (0.244, 0.263)	0.258 (0.249, 0.268)	0.263 (0.253, 0.271)	0.264 (0.255, 0.274)	0.267 (0.258, 0.276)	0.271 (0.262, 0.281)	0.276 (0.266, 0.285)
intron (non rpt)	0.005 (0.001, 0.009)	0.011 (0.008, 0.015)	0.017 (0.014, 0.021)	0.020 (0.016, 0.023)	0.024 (0.020, 0.027)	0.030 (0.026, 0.033)	0.036 (0.032, 0.039)
5' IG ≤ 5kb	0.071 (0.065, 0.077)	0.073 (0.067, 0.079)	0.075 (0.069, 0.081)	0.076 (0.070, 0.082)	0.077 (0.071, 0.082)	0.079 (0.073, 0.085)	0.081 (0.075, 0.086)
5' IG > 5kb	0.029 (0.025, 0.033)	0.035 (0.031, 0.039)	0.041 (0.037, 0.045)	0.043 (0.039, 0.047)	0.046 (0.042, 0.051)	0.052 (0.048, 0.056)	0.057 (0.053, 0.062)
3' IG ≤ 5kb	0.015 (0.009, 0.021)	0.018 (0.012, 0.025)	0.022 (0.016, 0.027)	0.023 (0.017, 0.029)	0.025 (0.019, 0.031)	0.028 (0.022, 0.034)	0.031 (0.025, 0.037)
3' IG > 5kb	0.039 (0.035, 0.043)	0.044 (0.040, 0.049)	0.049 (0.045, 0.054)	0.052 (0.047, 0.056)	0.055 (0.050, 0.059)	0.060 (0.055, 0.064)	0.064 (0.059, 0.069)

Constraint estimates for all genes, including AS and ST genes, are estimated assuming different equilibrium GC content values in the range of 25-50%.

Table 2.7. Selective constraint estimates for hominid sequence types with equilibrium GC content, corrected for the local recombination rate and GC content.

Sequence type	C
zero-fold	0.735 (0.723, 0.746)
four-fold	0.245 (0.216, 0.270)
5'UTR	0.169 (0.143, 0.197)
3'UTR	0.192 (0.178, 0.206)
intron	0.036 (0.032, 0.040)
5' IG ≤ 5kb	0.094 (0.083, 0.105)
3' IG > 5kb	0.039 (0.033, 0.045)
5' IG ≤ 5kb	0.091 (0.080, 0.101)
3' IG > 5kb	0.038 (0.032, 0.044)

Selective constraint estimates after correcting the assumed equilibrium GC content for the regional recombination rate and GC content per 1Mb segments (correction was done based on Duret and Arndt 2008)

It is also likely that local variation in the equilibrium GC content does not substantially affect my constraint estimates for murids, since the recombination rate is lower and less varied in mouse than in human (Jensen-Seaman et al. 2004) and the distribution of isochores is very similar in human and mouse (Costantini et al. 2009).

To test whether differences in constraint are caused by uncertainties in the assumed equilibrium GC content and consequent mis-assignments of directional changes (i.e. AT→GC and GC→AT), I estimated constraint based on the two pairwise rates A↔T and G↔C (see Table 2.8). In hominids, the results do not differ significantly from my previous analysis assuming an equilibrium GC content of 0.37, with the exception of two sequence types. Constraint estimates are significantly lower, when calculated based on the pairwise rates only, for 5'UTRs and for proximal 5' intergenic

regions (by 0.105 and 0.037, respectively). For murids, constraint estimates are lower for zero-fold, four-fold and 5' intergenic regions (by 0.011, 0.153 and 0.026, respectively), while slightly higher for 3'UTRs (by 0.026). However, with the exception of 5' intergenic regions the directions of the differences are the same. In this case, estimates become non-significantly different for proximal 5' intergenic regions (higher in hominids) and become higher for murids at distal 5' intergenic regions (non-significantly different).

Table 2.8. Pairwise-rate dependent selective constraint estimates on hominid and murid sequence types.

Sequence type	C hominid	C murid
zero-fold	0.724 (0.706, 0.740)	0.791 (0.783, 0.799)
four-fold	0.210 (0.148, 0.267)	-0.041 (-0.065, -0.018)
5'UTR	0.063 (0.000, 0.123)	0.462 (0.443, 0.481)
3'UTR	0.204 (0.179, 0.226)	0.303 (0.291, 0.315)
intron	0.042 (0.035, 0.049)	0.029 (0.024, 0.034)
5' IG ≤ 5kb	0.058 (0.039, 0.076)	0.061 (0.052, 0.070)
3' IG > 5kb	0.037 (0.028, 0.046)	0.055 (0.049, 0.061)
5' IG ≤ 5kb	0.080 (0.061, 0.099)	0.014 (0.004, 0.023)
3' IG > 5kb	0.039 (0.029, 0.049)	0.061 (0.054, 0.068)

Constraint was estimated based on A↔T and G↔C substitutions.

The results using A↔T and G↔C changes are in general agreement with the results described for four-fold sites and 5'UTRs (for four-fold sites constraint is estimated to be: 0.210 and -0.041, whereas for 5'UTRs it is 0.063 and 0.462, for hominids and murids, respectively).

Changing the assumed equilibrium GC content and calculating constraint based on pairwise rates both have some effect on my estimated constraint, but these do not change the trends I previously described. First, constraint is significantly positive in intergenic regions and in introns. Second, constraint on 5'UTRs is significantly

higher in murids than in hominids. Finally, constraint at 4-fold sites is significantly higher in hominids than in murids (see Table 2.4 to 2.8).

2.4.4 Patterns of constraint in flanking intergenic regions

I plotted intergenic constraint in hominids and murids against distance from the transcription start or end position using data pooled over ST and AS genes (Figure 2.5). The plots suggest that, on average, intergenic regions at the 5' and 3' flanks of hominid genes are under significant selective constraint. Constraint is significant even as far as 15-20 kb from the transcription start and end positions, but drops somewhat in deeper intergenic regions. In murids, however, several differences are apparent. First, in murids there is a sharp increase in constraint within ~1kb to 5' end of genes (as also observed by Keightley & Gaffney 2003 and Gaffney & Keightley 2006, which contrasts with the moderate increase in constraint over a much longer distance (over 15kb) of 5' end of hominid genes. Second, mean constraint for the 5kb 5' or 3' of genes is slightly higher in hominids than in murids, but this trend disappears deeper into intergenic regions where the level of constraint is higher in murid sequences.

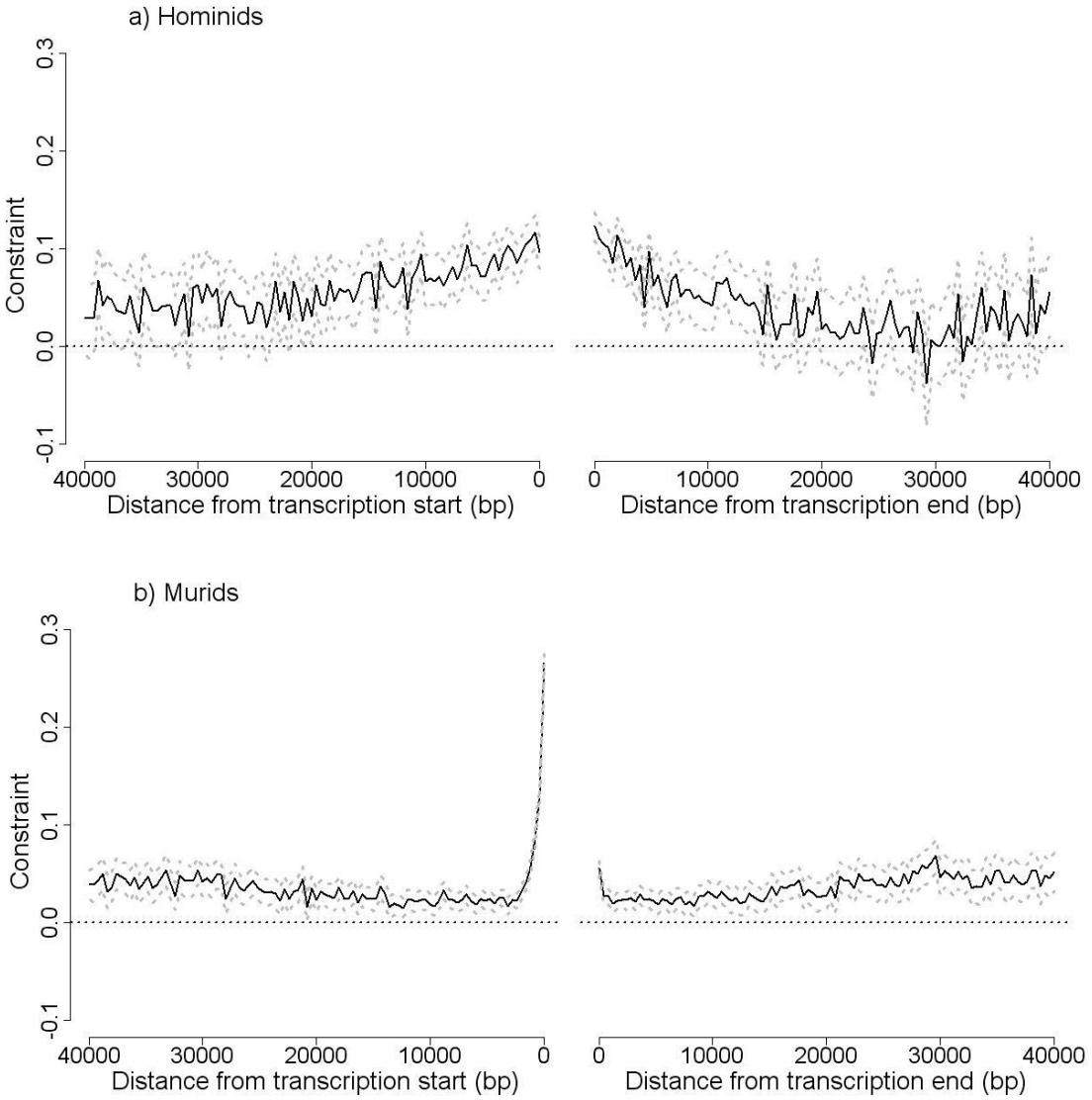

Figure 2.5. Intergenic constraint as a function of distance from the transcription start and end position of genes in hominids (a) and murids (b). Constraint estimates are calculated for non-CpG-prone sites. 95% confidence intervals are shown by dashed lines.

2.5 Discussion

In this study, I have analysed variation in selective constraint for different sequence types associated with single transcript (ST) and alternatively spliced (AS) genes in hominids and murids, and identified unexpected differences between the taxa.

2.5.1 Comparisons of constraint in flanking regions of genes between hominids and murids

A previous study of selective constraint in intergenic regions flanking protein-coding genes of hominids inferred that selective constraint was nearly absent in the 5' region, and was around 0.07 in the 3' region (Keightley et al. 2005). However, in the corresponding regions of murids, constraint was estimated to be moderately strong, with values of 0.17 and 0.19, respectively. The regions analysed include 5' and 3'UTRs, which are known to contain many regulatory sites (Shabalina & Spiridonov 2004; Hughes 2006; Chatterjee & Pal 2009), along with fragments of flanking intergenic regions of variable length, containing the transcription start sites (Dermitzakis & Clark 2002) and the promoter region (Frith 2006). In contrast, in my study I found moderate constraint on these regions in hominids, with values exceeding 0.15 for 5' and 3' UTRs, and values of around 0.10 for proximal intergenic regions. Furthermore, my sliding window analysis of the change in constraint from the transcriptional start/end positions towards deep intergenic regions also suggests that these regions are under constraint in hominids. There are likely to be several explanations for the differences between the two studies. The most important of these are: (a) the neutral standards used (non-first introns previously and ARs here), (b) the sequence types included in each analysis (5' and 3' UTRs and their introns along with IG regions previously, whereas only IG regions were considered here), and (c) the sampling of genes (1,000 genes were analysed previously, and over 15,000 genes are analysed in the present study).

2.5.2 Differences in selective constraints on non-coding DNA between hominids and murids

In most of the cases constraint estimates are higher in murids than in hominids (i.e. at distal 5' and 3' IGs, 5' and 3' UTRs and zero-fold sites). This agrees with a prediction of the nearly neutral theory of Kimura and Ohta, i.e. that selective constraint is expected to be lower in populations with low effective population sizes (N_e) (Kimura 1983; Ohta & Gillespie 1996), but it is important to note that mutational biases mitigate the effect of selection especially in the case of nearly neutral mutations (when $|N_e s|<1$) (see McVean & Charlesworth 1999; Takano-Shimizu 1999).

My results presented here reveal unexpected differences between hominids and murids (Figure 2.6), the most striking of which are: (a) the 2.9 times higher constraint on 5'UTRs in murids relative to hominids (0.48, 0.17, respectively), (b) higher constraints on proximal intergenic regions in hominids, which is most pronounced at the 3' end of genes, and (c) substantially higher constraint at four-fold sites in hominids than in murids (0.25, 0.11, respectively).

Mean constraint on 5'UTRs is approximately three-fold higher in murids than in hominids. This difference will have contributed to the difference in constraint in flanking regions between hominids and murids observed previously (Keightley et al. 2005), which included UTRs. However, I detect only a very small difference in constraint on flanking 5' intergenic regions (<0.02). Any genomic degradation in hominids therefore only appears to affect the 5'UTRs of genes, and is not a general phenomenon in the genome. The available annotations of 5'UTRs are 20% longer in hominids than in murids (Table 2.1). When upstream open reading frames and 5'UTRs that have overlapping annotations with introns are considered, this difference becomes 40%, which is comparable to a previous estimate of 37% for a smaller dataset (Vinogradov & Anatskaya 2007).

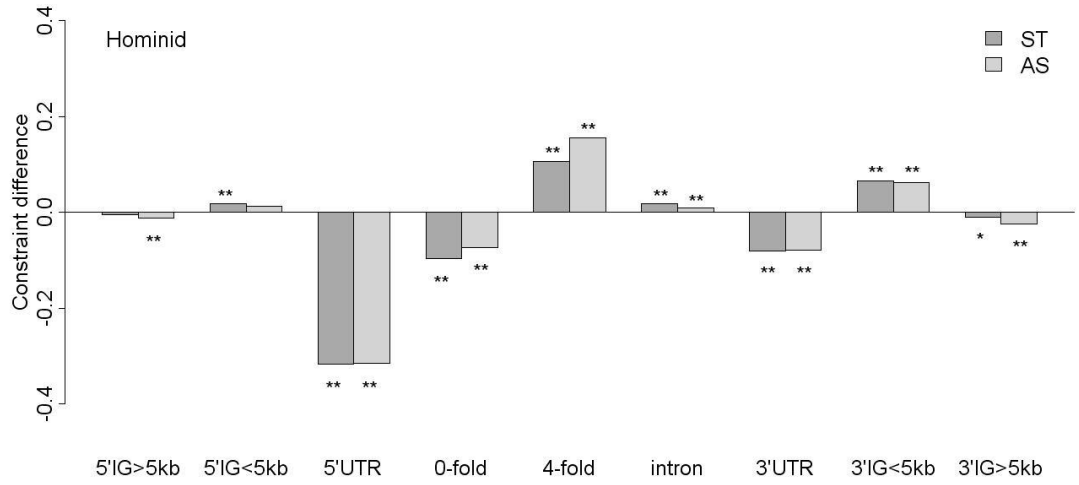

Figure 2.6. Constraint difference between different sequence categories of hominids and the corresponding categories in murids in single transcript (ST) and alternatively spliced (AS) genes, at non-CpG-prone sites. Asterisks indicate significant differences in constraint between hominids and murids (* P<0.05, ** P<0.01).

It has been suggested that the length of 5'UTRs is entirely driven by stochastic mutational processes (Lynch et al. 2005). However, a weak positive correlation between length and GC content has been found, and 5'UTRs are not as long in GC rich regions, which are unexpected under a neutral model (Reuter et al. 2008). It has been proposed that longer 5'UTRs in humans than mice are a consequence of more complex regulation at the translational level in humans (Vinogradov & Anatskaya 2007). The fact that 5'UTRs influence mRNA stability, regulate translation by providing internal ribosome entry sites and binding sites for trans-acting factors (Hughes 2006; Chatterjee & Pal 2009) and may interact with microRNAs (Lytle et al. 2007) is strong evidence that these sites evolve under selection. It is therefore surprising that I find evidence for low constraint on hominid 5'UTRs, supposedly with a higher level of transcriptional regulation, relative to murid (an average of 0.17 compared to 0.48, respectively). One possible explanation may be a difference in the accuracy of annotation between hominids and murids, i.e. that murid 5'UTRs contain

a higher concentration of unannotated open reading frames. However, a simple calculation suggests that this is quite unlikely, since approximately 50% of the sites within murid 5'UTRs would need to belong to unannotated open reading frames to generate comparable constraint to that observed for hominids. Another explanation might be a higher rate of adaptive substitutions in hominids, although this also seems to be unlikely, since positive selection is more likely to be effective in 5'UTRs of murids than hominids (Keightley et al. 2005; Eyre-Walker & Keightley 2009). The lower level of constraint and longer 5'UTRs of hominids may therefore be a consequence of less effective selection due to lower effective population size , but it remains an open question as to why this would have such a pronounced effect on 5'UTRs.

Mean constraint estimates for proximal 5' intergenic regions are higher for hominids than murids, but the difference is small (constraint is 0.095 and 0.077 respectively), and there is a wider margin at the 3' end (0.089 and 0.025) which is virtually unaffected by the assumed equilibrium GC content (Table 2.6). The observed higher constraints in proximal intergenic regions of hominids are unexpected, on the predictions of the nearly neutral theory. One possible explanation is that IG regions of hominids contain constrained, unannotated UTR sequences. It is known, for example, that there are usually multiple transcription start sites that are frequently loosely defined (Frith 2006) and usually weakly conserved in hominids (Birney et al. 2007), so if transcription start sites are better annotated in murids, then this may cause some of the differences in constraint observed in the 5' regions. However, the extent of lack of annotation is unknown, and this may equally well affect my constraint estimates for murid flanking regions.

A process that can also cause differences in constraint between taxa is the evolutionary turnover of functional sites. While protein-coding genes under strong selective constraint remain relatively invariant over long evolutionary periods, short stretches of functional non-coding DNA, such as transcription factor binding sites or transcription start sites, have been demonstrated to undergo evolutionary turnover

(Frith 2006; Moses et al. 2006). If new elements arising by turnover are preferentially located within lineage specific sequences (e.g., in new copies of transposable elements, see Pereira et al. 2009), then this may cause differences in the level of constraint between taxonomic groups, especially between taxa like hominids and murids, that differ in their divergence times. This may also explain the differences I see in the fine scale comparison of constraint in intergenic regions towards transcription start and end positions of genes (Figure 2.5). Lower constraint in intergenic regions over ~1kb from transcription start/end positions in murids relative to hominids may therefore be a consequence of turnover. Higher constraint in murids close to transcription start sites may be caused by functional sites with slow turnover, or by unspecified transcription start sites and so by the inclusion of 5'UTR sequences. While turnover may explain some of the differences in constraint estimates between hominids and murids, the fact that constraint estimates based on human/chimp or human/macaque (see next chapter) are similar or, even higher in the latter comparison, suggests that the overall effect of turnover is low.

The most intriguing observation is the two-fold higher constraint at four-fold sites in hominids relative to murids, which raises the question as to what selective processes can explain the observed difference? First, while in mammals selection on translational efficiency is generally considered to be weak (Chamary et al. 2006), it has been suggested that selection operates for translational accuracy in human and mice and the strength of association between preferred codons and conserved amino acids are similar in the two species (Drummond & Wilke 2008), making it unlikely that these factors would explain the difference. Second, selection may also operate for optimal mRNA stability and structure (see Chamary et al. 2006) which may yield differences in constraint between hominids and murids, if human mRNAs fold into more complex structures. At present, in the absence of experimentally determined mRNA structures, studies of selection on mRNA structure are usually based on in silico structure predictions (Chamary & Hurst 2005), which do not necessarily reflect the in vivo structure of mRNAs, so that estimates on the strength of selection may be

biased. Third, significantly higher constraint at four-fold sites in hominid AS genes relative to ST genes suggests that splicing and splicing regulation may constrain evolution at these sites, although constraint is not significantly different between AS and ST genes in murids. Indeed, it is known that proper splicing requires the presence of exonic splice enhancer (ESE) and silencer (ESS) sequences (Blencowe 2000), which are known to be under selective constraint (Parmley et al. 2007). Protein sequences of longer-lived taxa, such as hominids, may contain more ESE and ESS sequences, or these sites may operate under stronger selection in hominids in order to maintain proper splicing throughout the individual's lifespan. Last, the low level of constraint on hominid 5'UTRs, coupled with the high level of constraint on four-fold sites and on proximal intergenic regions relative to murids, can also be taken as evidence for reorganisation of functional sites between the two taxa, and may suggest regulatory differences at the level of translation.

It is well established that in several organisms mutations at four-fold sites are selected against, due to selection on codon usage, (Rocha 2006; Drummond & Wilke 2008), mRNA structure (Chamary & Hurst 2005) or mRNA splicing (Parmley et al. 2007); as a consequence, the d_N/d_S ratio, which has been frequently used to detect the strength and direction of selection (e.g., Dorus et al. 2004; Wang et al. 2008), may lead to false positives in the detection of adaptive evolution. My finding of higher four-fold constraint in hominids suggests that this bias more strongly affects hominid estimates and may well exceed 20%.

2.5.3 Genomic selective constraint and the deleterious mutation rate in hominids

My results show that AS genes are more than two times longer, on average, than ST genes, primarily due to the presence of longer introns in alternatively spliced genes, and to a lesser extent longer protein-coding and UTR sequences. The mean constraint levels on sequences associated with AS genes are also higher than those observed for ST genes. Thus, previous estimates of genomic selective constraint (C), based on ST

genes only, and estimates of the deleterious mutation rate (U) derived from this, are likely to be downwardly biased.

The estimated constraint values and the known overall length of the different sequence types allow us to estimate the genomic contribution of each type to the total number of constrained sites in the human genome (Table 2.4). I estimate that there are 78.8 Mb of constrained sites in the non-repeat fraction of the hominid genome, 17% and 39% coming from ST and AS genes, respectively, and the remaining 44% coming from intergenic regions. A larger contribution from complex regions is attributable to three factors, namely the larger number of genes in this category, the larger number of nucleotides associated with their sequence types, and the higher mean level of constraint in complex categories.

Summing over constrained sites in coding and non-coding regions, I estimated that protein-coding categories contribute 18.1 Mb of constrained sites to the total, and that non-coding sites contribute a total of 60.7 Mb (i.e. more than three times as many as protein-coding sites). Dividing the total number of constrained sites by the 1454.2 Mb of sites associated with the known, non-repetitive fraction of the genome gives an estimate of 0.054 for the average constraint per nucleotide in the hominid genome, a value that is in reasonable agreement with previous estimates in the range of 0.03 and 0.08 of hominid genomic selective constraints (Chiaromonte et al. 2003; Siepel et al. 2005; Birney et al. 2007). My genomic constraint estimate for murids, taking into account the non-equilibrium GC content in ARs, is 0.055, which is lower than a previous estimate of 0.087 (Gaffney & Keightley 2006), and is essentially the same as the value for hominids. Under the prediction of the nearly neutral theory I would expect genomic selective constraint to be lower in hominids than in murids, but since around 99% of the genome is non-protein-coding, genomic constraint is strongly dependent on the estimated constraint on introns and intergenic regions. If for example the equilibrium GC content in murids were 0.45 instead of 0.40 then the genomic selective constraint would be 0.060, which is around 10% higher than what

I estimated for hominids. This might be the case, since the equilibrium GC content in murids may be somewhat higher than in hominids (Khelifi et al. 2006).

I used the observed substitution rates in intronic and intergenic ancestral repeats to estimate the mutation rate per generation in genic and intergenic sequences (Table 2.9 and 2.10), on the assumption that transposable elements are neutrally evolving.

If the human and chimpanzee split occurred 6 Mya (Patterson et al. 2006), and the average generation time is 25 years (Eyre-Walker & Keightley 1999), then the estimated total number of mutations (M) in the repeat-free genome (exluding CpG dinucleotides) is 65 mutations per diploid genome per generation. CpG dinucleotides show 8-18 fold higher rates of evolution (Arndt et al. 2003; Lunter & Hein 2004; CSAC 2005) than non-CpG sites. Assuming CpG sites were 10-fold more mutable would contribute an additional 12 mutations, leading to a total of 77 mutations per diploid genome per generation. Multiplying M by the genomic selective constraint (C), I estimate the deleterious mutation rate per diploid per generation, U, to be 4.4 (see Table 2.10). The contribution to U from amino acid changing mutations is 0.8, whereas there are more than four times as many (3.4) deleterious mutations in non-coding sequences. That eighty percent of selectively constrained sites are located outside protein-coding sequences highlights the importance of extensive empirical studies, such as the ENCODE project (Birney et al. 2007), which aim to systematically identify functional elements within non-coding regions.

Table 2.9. Number of mutations per generation in genic and intergenic regions of single transcript and alternatively spliced genes in the haploid genomes of human and mice.

A. Human

ST	*N* site (Mb)	$f_{non\ CpG}$	d_{nCpG} (95% CI)	μ	$N_{non\text{-}CpG}$	N_{CpG}
Genic	202.9	97.73%	0.0103 (0.0102, 0.0105)	2.15×10^{-8}	4.26	0.99
Intergenic	442.1	98.35%	0.0117 (0.0116, 0.0119)	2.44×10^{-8}	10.62	1.78
Total					14.88	2.77

AS	*N* site (Mb)	$f_{non\ CpG}$	d_{nCpG} (95% CI)	μ	$N_{non\text{-}CpG}$	N_{CpG}
Genic	432.1	97.89%	0.0101 (0.1000, 0.0102)	2.11×10^{-8}	8.92	1.92
Intergenic	374.0	98.39%	0.0114 (0.0113, 0.0115)	2.37×10^{-8}	8.71	1.42
Total					17.63	3.34

B. Mice

ST	*N* site (Mb)	$f_{non\ CpG}$	d_{nCpG} (95% CI)	μ	$N_{non\text{-}CpG}$	N_{CpG}
Genic	256.7	97.72%	0.1570 (0.1559, 0.1581)	3.02×10^{-9}	0.76	0.18
Intergenic	552.0	98.47%	0.1677 (0.1663, 0.1690)	3.23×10^{-9}	1.75	0.27
Total					2.51	0.45

AS	*N* site (Mb)	$f_{non\ CpG}$	d_{nCpG} (95% CI)	μ	$N_{non\text{-}CpG}$	N_{CpG}
Genic	370.9	97.91%	0.1567 (0.1557, 0.1577)	3.01×10^{-9}	1.09	0.23
Intergenic	288.5	98.39%	0.1659 (0.1642, 0.1677)	3.19×10^{-9}	0.91	0.15
Total					2.00	0.38

$N site$ is the number of non repetitive nucleotides in genic/intergenic regions per single transcript (ST) and alternatively spliced (AS) genes; $f_{non\text{-}CpG}$ is the percentage of sites that are not part of a CpG dinucleotide, estimated based on the human and mouse genomic sequences; d_{nCpG} is the non-CpG substitution rate observed in genic and intergenic ARs, used here to estimate of the neutral mutation rate; μ is the neutral mutation rate per generation. $N_{non\text{-}CpG}$ is the number of mutations occurring in the repeat free human genome per generation at non CpG sites; N_{CpG} is the number of CpG mutations in the repeat free human genome per generation assuming 10 fold hypermutatbility; μ is calculated by assuming that human and chimpanzee and mouse and rat diverged 6Mya and 13Mya, and generation times of hominids and murids are 25 years and 0.5 years, respectively.

Table 2.10. Comparison of constraint, number of mutations per generation and the deleterious mutation rate between hominids and murids.

	C	$M_{non\text{-}CpG}$	M_{CpG}	$U_{non\text{-}CpG}$	U_{CpG}	U_{total}
Hominid	0.057	65.0	12.2	3.7	0.70	4.4
Murid	0.056	9.02	1.66	0.51	0.094	0.60

C is the genomic selective constraint per nt, M is the total number of mutations per diploid per generation and $U_{non\text{-}CpG}$ and U_{CpG} are the number of deleterious mutation at non CpG and at CpG sites per diploid per generation, assuming 10-fold CpG hypermutability. M is calculated by assuming divergence times of 6 My and 13 My and generation times of 25 years and 0.5 years for hominids and murids, respectively.

Although the estimates of the human neutral mutation rate per generation well agree with previous direct (Kondrashov 2003; Lynch 2010) and indirect (Nachman & Crowell 2000) estimates, the estimate of U is still strongly affected by both the divergence time and the generation time. For human/chimpanzee, divergence time estimates are between 4.1 and 7 Mya (Hobolth et al. 2007; Vignaud et al. 2002), and although hominids are known to have long generation times, a value of 20 years may be a better estimate than the 25 years previously assumed (Eyre-Walker & Keightley 1999; Nachman & Crowell 2000, but see Elango et al. 2006). Accounting for these uncertainties, the range for the deleterious mutation rate in hominids is 3.0-6.5. There are other reasons why my estimate of U might be different from the true value. First, I assume that the fraction of changes caused by adaptive evolution is negligible in mammals. In general, adaptive evolution would downwardly bias my estimate of constraint, although adaptive evolution in ARs would inflate the estimate of the mutation rate and lead to an upward bias. Second, if intergenic and intronic ARs contain substantial number of functional sites that are under purifying selection, then my constraint and neutral rate estimate would both be downwardly biased and I would consequently underestimate U. Indeed, there is evidence that some

transposable elements have become functional and are subject to selection (Kamal et al. 2006; Lowe et al. 2007; Pereira et al. 2009), although the fraction of functional elements is thought to be less than 0.1% (Lunter et al. 2006; Lowe et al. 2007). Third, in my analysis I have not accounted for the contribution of RNA genes to the total number of constrained sites, but at present, annotated RNA genes make up less than 0.01% of the genome. Fourth, I did not attempt to estimate selective constraint for tandem and microsatellite repeats, because sequencing and alignment of these regions is uncertain (Ponting 2008). Fifth, many insertion and deletion events are under purifying selection if they disrupt a functional sequence (Lunter et al. 2006) and by not considering these events I will have underestimated the deleterious mutation rate.

Throughout my analysis I have attempted to exclude mutations that result from spontaneous deamination of methylated CpG dinucleotides, using parsimony (i.e., by excluding sites preceded by nucleotide C or followed by G) (Meunier & Duret 2004; Khelifi et al. 2006). This is necessary because the inclusion of these sites may cause bias in constraint estimates (Keightley & Gaffney 2003; Gaffney & Keightley 2006). In the future, I anticipate that multi-species comparisons and advanced methods for constraint calculations, based on improved substitution models (e.g., Siepel & Haussler 2004; Duret & Arndt 2008), will help elucidate the fraction of constrained sites and the neutral rate of evolution in parts of the genome strongly affected by CpG hypermutability.

The high estimate of U has important implications for hominid evolution. Even my lower bound estimate of 3.0 deleterious mutations would lead to a mutational load (L) of 95% (i.e., the fraction of individuals that fail to contribute to the next generation) in hominids, assuming that fitness effects are multiplicative, so that the mutational load is $L = 1 - e^{-U}$ (Kimura & Maruyama 1966). Even if selection occurred in the germline or during the early prenatal stage of development (Crow 2000), it is difficult to envisage how such a high load could be tolerated by hominid populations, which have very low reproductive rates. Load could be reduced if

mutations have synergystic epistatic interactions on fitness, leading to non-independent elimination of mutations (Crow 2000).

It has been suggested that a relaxation in the strength of selection may lead to the accumulation of very slightly deleterious mutations in the human genome (Kondrashov 1995). Although this could in theory lead to a long term decline of fitness of human populations (Lynch 2010), it can be argued that this is unlikely to be the case for two reasons. First, in spite of the historically low effective population sizes of humans, the recent population size is continuously expanding and thought to be much larger than the long-term effective size. Second, although living conditions have improved during the last centuries, this is a short period relative to the evolutionary time scale. Under such circumstances, the long term effectiveness of selection, which depends on the product of Ne and s, is expected to be stronger. Finally, any degradation in fitness critically depends on the distribution of effects of new mutations in both coding and non-coding DNA. Evidence suggests that many non-synonymous mutations are effectively selected against (Keightley & Eyre-Walker 2007; Eyre-Walker & Keightley 2007; Boyko et al. 2008). However, the distribution of selective effects of mutations in non-coding DNA is still essentially unknown.

3 Variation in selective constraint at synonymous and non-synonymous sites among vertebrates

The work described in this chapter is submitted to MBE (Eory, L. Variation in selective constraint at synonymous and non-synonymous sites in vertebrates). This chapter has the same format as the published paper and the text herein contains only slight modifications.

3.1 Abstract

Although synonymous sites have long been assumed to evolve neutrally, and are still the most widely used neutral standards for assessing the direction and strength of selection, there is some evidence that synonymous sites are under selective constraint, even in mammals. Here, I quantify the level of selective constraint on protein-coding sequences in different vertebrate species by comparing rates of evolution of synonymous and non-synonymous sites to that for ancestral transposable elements. These elements are among the best candidates for neutrally evolving regions of the genome. My results show that constraint at synonymous sites varies widely among mammalian species, suggesting that the strength of selection, when measured by the d_N/d_S ratio, may be overestimated by a factor of at least 11-67% on average, depending on the species of interest. This suggests that the d_N/d_S ratio may not be reliable for determining the strength of selection, and in cases when there is strong selection on synonymous sites, may lead to false positives in detecting adaptive evolution. Controlling for selection at synonymous sites is especially necessary when the d_N/d_S ratio is used to assess the differences in the mode and direction of selection between species. Testing the results of the d_N/d_S ratio by choosing neutral standards other than synonymous sites may therefore be necessary if the properties of selection are to be properly understood. It is also important to note that not taking account of the effect of N_e and/or generation time on estimates of strength of selection may lead to spurious correlation between strength of selection and divergence.

3.2 Introduction

While the majority of mutations that occur in vertebrate genomes are likely to behave as selectively neutral and have no functional consequences, some fraction of them have moderate or strong effects on the fitness of an individual. Beneficial mutations have a higher probability of fixation than neutral mutations, while those with harmful effects are expected to be removed by purifying selection (Kimura 1983). The effect of positive selection is thought to be causal in the emergence of new traits that lead to new morphological and physiological characteristics and thereby drives the evolution of species (Darwin 1859). Therefore the extent to which selection, especially positive selection, affects vertebrate genomes has been the subject of many studies (Bakewell et al. 2007; Kosiol et al. 2008; Wolf et al. 2009; Pollard et al. 2010).

Most of our knowledge concerning the direction and strength of selection acting in nucleotide sequences is based on the simple assumption that, in the case of neutrality, the rate of change per site should be proportional to the mutation rate per site (Kimura 1983). Deviations from this relationship are considered to be the results of selection. For historical and practical reasons, synonymous sites, mostly nucleotides at third codon positions where a mutation does not cause a change in the protein sequence itself, are considered to be evolving nearly neutrally, and have frequently served as neutral standards in methods employed to test for selection (e.g. Keightley & Eyre-Walker 2000; Bakewell et al. 2007; Kosiol et al. 2008). Indeed, the rate of evolution at synonymous sites is not only used to infer the phylogeny of species (Miller et al. 2007), but also used to estimate the neutral evolutionary rate. It has been employed for this purpose in tests such as the McDonald-Kreitman test (McDonald & Kreitman 1991), which can indicate the extent of adaptive evolution (Eyre-Walker 2006), and in the d_N/d_S ratio, which is frequently used to test the direction and strength of natural selection (Hurst 2002).

Although it has frequently been assumed that synonymous sites evolve neutrally, especially in mammalian species that have small effective population sizes, there is evidence that synonymous sites are under some selective constraints (Chamary et al. 2006; Drummond & Wilke 2008; Pollard et al. 2010, see also Chapter 2). If constraint at these sites is substantial and varies among species, then this could affect phylogenies inferred based on synonymous sites and in turn any model of selection based on such phylogenies (Yang & Nielsen 2000; Pollard et al. 2010). In cases when synonymous sites are under strong constraint, the d_N/d_S ratio may falsely suggest adaptive evolution (i.e., $d_N/d_S>1$) (Chamary et al. 2006; Wolf et al. 2009), and interspecies comparisons (e.g. Bakewell et al. 2007) may be unreliable if between species variation in synonymous constraint is unaccounted for.

In this study, I aim to quantify synonymous and non-synonymous selective constraints in eighteen vertebrate species, based on comparative genome analysis of closely related species pairs. I explore the variation in constraint between species, and test the effect of factors, such as effective population size (N_e) and evolutionary distance on constraint estimates as it was found that these are positively correlated with the d_N/d_S ratio (Rocha et al. 2006; Ellegren 2009; Wolf et al. 2009). My results suggest that not accounting for the effect of N_e and/or generation time (GT) may cause spurious correlations between constraint (the d_N/d_S ratio) and divergence. I go on to discuss the consequences of constraint at synonymous sites on the results of applying methods to quantify selection. We previously inferred that selective constraint at synonymous sites is higher in hominids than in murids (Chapter 2) which is at odds with a prediction of the nearly neutral theory of molecular evolution (Ohta 1995). Here I test whether the two fold difference we previously observed can be explained by a change in the level of selection in the hominid or in the murid lineages, respectively.

Selective constraint is defined as the fraction of mutations that are deleterious and therefore are removed by purifying selection (Keightley & Gaffney 2003 and Materials and Methods), and is in many respects analogous to 1- d_N/d_S. The main

differences are 1) the use of orthologous transposable elements (so called ancestral repeats – ARs) as a neutral standard to estimate the mutation rate and 2) the assumption that mutation rates vary in vertebrates on a scale of 1Mb (Gaffney & Keightley 2005; Duret & Arndt 2008). I assume that ARs evolve close to neutrality, since there is evidence that transposable elements evolve in a largely neutral fashion (Lunter et al. 2006) and contain the lowest fraction of conserved nucleotides compared to other types of sequence (Pollard et al. 2010).

3.3 Materials and Methods

Data. Genomic sequence data and annotation for eleven, closely related (average divergence in ancestral repeats is less than 0.3) species pairs were accessed and downloaded from the Ensembl MySQL server using the Perl API (Hubbard et al. 2009) (see Table 3.1, Figure 3.1). I chose BLASTZ (Schwartz et al. 2003), EPO or EPO-LOW-COVERAGE (Hubbard et al. 2009) whole genome alignments, in order of preference, to obtain alignments of the protein-coding genes for the species pairs, and realigned them using MAVID (Bray & Pachter 2004). One species from the species pair, preferably the one with better sequence coverage and/or more reliable gene annotation was chosen as the reference species, for which gene annotations were downloaded from Ensembl (release 57). These annotations were mapped onto the genome of the other species (referred to as the target genome) using the fact that gene structure in mammals is well conserved (Roy et al. 2003). All genes that fulfilled the system of criteria for orthology described below were analysed.

I assumed that any transcript was orthologous between the reference and target species if the target sequence started with the same codon as the reference, or both started in a start codon. Transcripts were also checked to ensure they ended in a stop codon, and did not contain premature stop codons or frameshift-mutations. Genes were considered to be valid if they contained at least one valid orthologous transcript in the target genome.

Transposable element (TE) annotations were downloaded from Ensembl, and those TEs with a putatively orthologous sequence in the target genome (ancestral repeats - ARs) were used as neutral standards, by pooling them together over 1Mb segments of the genomes. I only used intronic ARs in order to account for the likely effects of transcription coupled mutation and repair (Majewski 2003; Green et al. 2003) on substitution rate estimates. Low complexity regions, tandem and microsatellite repeats were masked off from the analysis, since the sequencing and alignment of these regions may be problematic.

To reduce possible biases in the estimates due to low quality or misaligned sequences, a simple masking protocol was employed to exclude sites that were likely to be non-orthologous between the reference and target genomes. Divergence was calculated in sliding windows of 40 alignment columns, and multiple hits were corrected by Kimura's two-parameter method (Kimura 1980). Any region covered by 50 or more contiguous windows within which divergence was abnormally high (higher than three SD from the mean divergence) or the window contained less than 50% valid aligned bases (due to gaps in the alignment), was masked out of the alignments.

Data analysis. Substitution rates were estimated at all sites and at non-CpG-prone sites (sites not preceded by C or followed by G in either species). It was assumed that evolutionary rate at non-CpG-prone sites was unaffected by context dependent CpG hypermutability (Meunier & Duret 2004; Nikolaev et al. 2007) a typical feature of mammalian genomes (Bird 1980). For gene sequences, divergences at four-fold sites were calculated only for codons where both aligned codons code for the same amino acid, and at most a single nucleotide change had occurred. I assumed throughout this study that intronic ARs evolve free of evolutionary constraints, so that their evolutionary rates can be used to estimate the neutral divergence values for any sequence type. Calculation of constraint was done using the method of Halligan et al. (2004). To account for the differences in GC content between different sequence types, the method corrects for the difference in AT→GC and GC→AT mutation

rates. This method assumes that the directional rates depend on the equilibrium GC content. These rates are then used to estimate the expected number of substitutions (E) in an adjacent putatively functional sequence, here local ARs, by use of the equation $E=\sum_{i=1}^{4} k_i m_i$ where k_i is the rate of change and m_i is the corresponding number of nucleotide sites in the sequence of interest where a type of the four distinct changes (A↔T, G↔C, AT→GC and GC→AT, indexed with i) could have occurred. Then, by counting the number of substitutions (O) in the focal sequence, I calculate selective constraint as $C_{seq}=1-O_{seq}/E_{seq}$ (Halligan et al. 2004). Data were pooled together over 1Mb blocks, by starting the blocks with a valid gene and 95% confidence intervals for divergence and constraint were obtained by bootstrapping the dataset 1,000 times by block. The equilibrium GC content (p_e) was set to be 0.4 along chromosomes for all species-pairs, based on the work of Duret & Arndt (2008) on hominids. This is unlikely to hold for two reasons. The p_e of any sequence depends on the ratio of AT→GC and GC→AT mutation rates which is known to vary 1) along chromosomes (Duret & Arndt 2008) and 2) between species (Khelifi et al. 2006). While it was shown that regional variations do not have a large effect on the estimated constraint, uncertainties in p_e may cause bias in the estimates, see Chapter 2.

Effective population size (N_e) estimates were taken from the literature (Eyre-Walker et al. 2002; Yu et al. 2004; Piganeau & Eyre-Walker 2009; Mank et al. 2010). As the estimates vary from source to source and depend on several factors (e.g. quality of polymorphism data, estimates of mutation rate and divergence time, and many additional factors reviewed by Charlesworth 2009), generation time data were also collected to test the results based on N_e. When N_e or generation time (GT) estimates were available for both species in the species pair, then the average value was taken in the analysis.

Significance of the Pearson correlations and partial correlations between selective constraint, N_e, *GT* and divergence were calculated by resampling the datasets with Fisher-Yates shuffling (Fisher & Yates 1948).

To deal with a possible effect of non-independent data points in the Primate and Artiodactyla groups on the correlation estimates, correlations were also calculated on data from eight fully independent species pairs (leaving out the human-macaque and any two pairs from the Artiodactyla group).

3.4 Results and Discussion

I analysed the level of constraints at synonymous and non-synonymous sites in eleven closely related vertebrate species-pairs (see Table 3.1), which include ten mammalian and one avian species pair (for their phylogeny see Figure 3.1). Putatively orthologous genes were identified by applying a set of criteria described in the methods section, and a simple sequence filtering method was applied to remove nucleotides with poor sequence quality, bad alignments or unreliable annotations. The numbers of genes (and so the total number of sites in the analyses) fulfilling these requirements varied widely, being highest for hominids and murids, and depended on several factors, such as the quality of annotation, sequence coverage and alignment types (see Table 3.1 and Table 3.2). For example, the number of valid genes were generally the lowest when the target genome was of low coverage (~2x), regardless of the alignment type; also, in spite of their similar sequence coverage and known/all annotation ratios, I got a lower number of valid genes for the pig-cow pair as opposed to macaque-marmoset, most likely because macaque is more closely related to human, so that the human annotations used by Ensembl to project genes on to genomes are more likely to lead to proper annotations for macaque than for the more distantly related pig genome (see Table 3.1 and Schneider et al. 2009). The relatively low divergence between the species pairs analysed here enabled reliable substitution rate estimates to be obtained, while providing enough statistical power for the analysis.

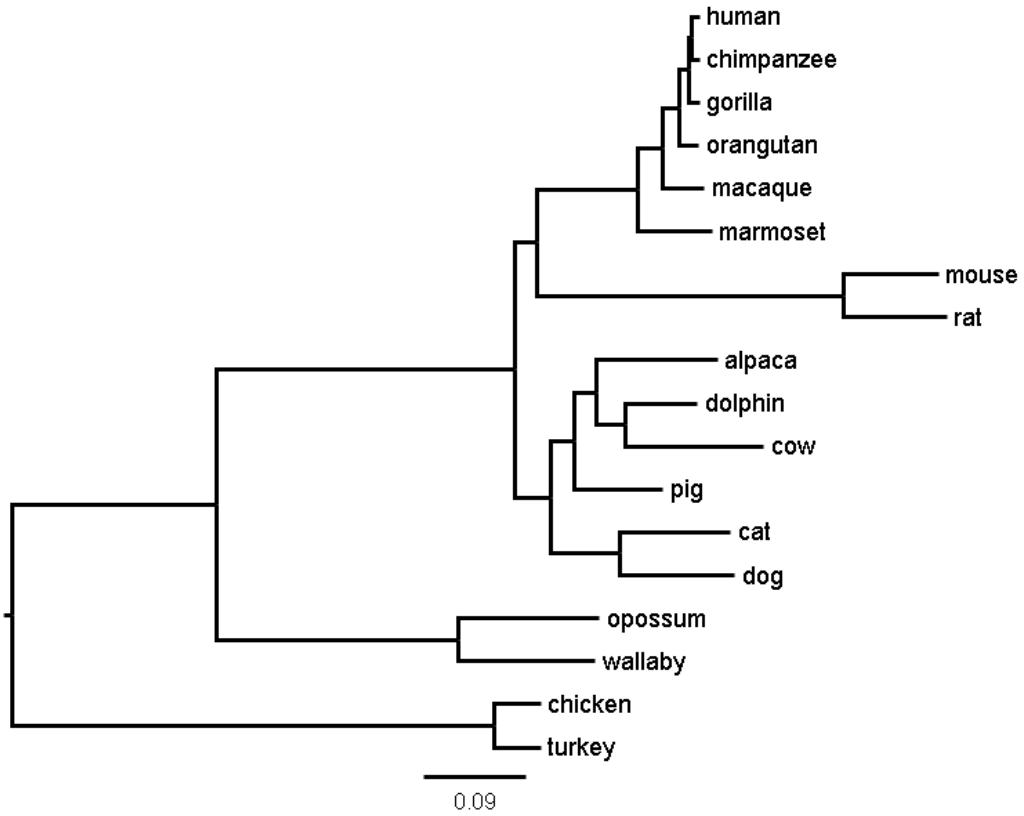

Figure 3.1. Phylogeny of the species analysed. Branch length estimates are based on Miller et al. (2007)), given for synonymous sites. Figure was generated by Andrew Rambaut's program FigTree v1.3.1.

3.4.1 Mutation rate and selective constraint estimates

Mutation rate estimates for ancestral repeats (ARs) were obtained at non-CpG-prone sites (e.g. Meunier & Duret 2004; Nikolaev et al. 2007; see Materials and Methods) in order to avoid possible biases caused by inflated evolutionary rates at CpG sites (S. I Nikolaev et al. 2007; A. P. Bird 1980). Accordingly, divergence rate estimates, which are known to be lower for non-CpG-prone sites in hominids and murids, are in general significantly lower in mammals and birds, suggesting a role for CpG hypermutability in these species. The only exception was observed for the opossum-

wallaby pair, for which divergence is slightly higher at non-CpG-prone sites (see Table 3.3). Using the substitution rates calculated for ARs, I estimated mean selective constraints at non-synonymous and synonymous sites (Table 3.4). Non-synonymous constraint is around 0.80, and the estimates are in the range of 0.71-0.87. The values are the lowest for the human-chimp and the highest for the chicken-turkey comparisons. Synonymous constraint is positive in all cases and varies more widely. Estimates are in the range of 0.10-0.39, with an average around 0.20, and the lowest estimate is for pig-cow and the highest for chicken-turkey. The mean non-synonymous and synonymous constraint estimates are in good agreement with the estimated constraint of 0.71 at second codon positions (non-synonymous) and of 0.25 at third codon positions (a mixture of synonymous and non-synonymous sites) in mammals, based on statistical phylogenetic tests (Pollard et al. 2010).

3.4.2 Effect of N_e on selective constraint

There are several factors, notably the effective population size (N_e) and evolutionary distance, which are known to covary with the d_N/d_S ratio and so these are expected to covary with the selective constraint estimates. Keightley & Eyre-Walker (2000) found a positive correlation between the genomic deleterious mutation rate and generation time in a wide range of animal taxa, the former of which depends on both selective constraint and the mutation rate. More specifically, Ellegren (2009) found a negative correlation between the d_N/d_S ratio and N_e in mammals, based on the d_N/d_S ratio estimates of Kosiol et al. (2008). According to the nearly neutral theory of molecular evolution (Ohta 1995), selection is efficient in purging deleterious mutations and driving advantageous mutations to fixation if the selection coefficient ($|s|$) is greater than $1/N_e$, while mutations with $|s|<<1/N_e$ are effectively neutral and their fate is largely determined by chance events.

Table 3.1. Summary of the data used in the analysis.

Abbrev	Ref species	Tar species	Ref coverage	Tar coverage	Known/All	Alignment	No valid genes
A	Human [1]	Chimpanzee [2]	Deep	7.0x	99%	BlastZ	16887
B	Human [1]	Macaque [3]	Deep	5.0x	99%	BlastZ	15695
C	Orangutan [4]	Gorilla [5]	6.0x	35.0x	19%	EPO	8680
D	Macaque [3]	Marmoset [6]	5.0x	7.0x	14%	EPO LOW	7668
E	Mouse [7]	Rat [8]	6.5x	7.0x	99%	BlastZ	15577
F	Cow [9]	Dolphin [10]	7.0x	2.0x	96%	EPO LOW	4067
G	Cow [9]	Alpaca [10]	7.0x	2.0x	96%	EPO LOW	1424
H	Pig [11]	Cow [9]	4.0x	7.0x	16%	BlastZ	5228
I	Dog [12]	Cat [13]	7.6x	2.0x	13%	EPO LOW	2032
J	Opossum [14]	Wallaby [10]	6.8x	2.0x	3%	BlastZ	2319
K	Chicken [15]	Turkey [16]	6.0x	30.0x	95%	BlastZ	6526

Gene annotations in the reference species (Ref species) were used to identify orthologous genes in the target species (Tar species). Genome sequence coverage for the reference and target sequences are shown in columns Ref coverage and Tar coverage, respectively. The ratio of the number of known genes to the total number of annotated genes (known + predicted based on human sequences) are shown in column Known/All, while alignment type and number of valid genes are shown in the last two columns, respectively. Genomic sequence data are from the following sources: 1. IHGSC 2004, 2. CSAC 2005, 3. RMGSAC 2007, 4. Orangutan Genome Sequencing Consortium unpublished, 5. Gorilla genome provided by the Welcome Trust Sanger Institute, unpublished 6. Marmoset Genome Sequencing Consortium unpublished, 7. IMGSC et al. 2002, 8. IRGSC et al. 2004, 9. Liu et al. 2009 10. Mammalian Genome Project unpublished, 11. Swine Genome Sequencing Consortium unpublished, 12. Lindblad-Toh et al. 2005, 13. Pontius et al. 2007, 14. Mikkelsen et al. 2007, 15. Hillier et al. 2004 ,16. Dalloul et al. 2010.c

Table 3.2. Number of sites analysed.

Species	Number of basepairs analysed (Mb)		
	Non-Synonymous	**Synonymous**	**AR**
human–chimpanzee	89.76	12.73	1625.21
human–macaque	83.07	11.88	1041.95
orangutan–gorilla	44.86	6.49	789.29
macaque–marmoset	39.22	5.46	317.12
mouse–rat	81.47	12.17	619.89
pig–cow	23.01	3.53	29.72
cow–alpaca	6.12	0.77	25.13
cow–dolphin	17.93	2.70	74.89
dog–cat	7.68	1.23	18.32
opossum–wallaby	9.00	1.27	15.39
chicken–turkey	34.97	4.49	39.68

Table 3.3. Pairwise divergence estimates at ancestral repeat sequences.

Species	Non-CpG-prone sites	All sites
human–chimpanzee	0.00955 (0.00947, 0.0964)	0.0114 (0.0113, 0.0115)
human–macaque	0.0541 (0.0539, 0.0543)	0.0581 (0.0578, 0.0584)
orangutan–gorilla	0.0281 (0.0279, 0.0282)	0.0340 (0.0338, 0.0342)
macaque–marmoset	0.108 (0.108, 0.109)	0.113 (0.112, 0.113)
mouse–rat	0.160 (0.159, 0.161)	0.164 (0.163, 0.165)
pig–cow	0.177 (0.176, 0.178)	0.180 (0.180, 0.181)
cow–alpaca	0.224 (0.222, 0.226)	0.226 (0.224, 0.229)
cow–dolphin	0.159 (0.158, 0.160)	0.166 (0.165, 0.167)
dog–cat	0.184 (0.182, 0.186)	0.199 (0.196, 0.202)
opossum–wallaby	0.290 (0.287, 0.292)	0.282 (0.280, 0.284)
chicken–turkey	0.110 (0.109, 0.110)	0.116 (0.115, 0.117)

Divergence rate estimates are given separately for all sites and for non-CpG sites. 95% confidence intervals are shown in parentheses.

This predicts a positive correlation between selective constraint and N_e in vertebrates. Indeed, as expected, I find a strong positive correlation between non-synonymous constraint and N_e (for $\log_{10}(N_e)$ r_p=0.82, P=0.0024, df=8) (see Figure 3.2 and Table 3.4), which is similar to the observation of Ellegren (2009), using synonymous sites as a neutral standard. Correlations remained quantitatively unchanged when data were constrained to include only independent species comparisons (i.e. either human/chimp or human/macaque, and only one from pig/cow, cow/alpaca and cow/dolphin were kept for the correlation study; see Table 3.5). Table 3.2A suggests that the constraint estimate for the human-chimpanzee pair is downwardly biased. Indeed, it has been suggested that around 20% of the observed genome-wide divergence between the two species may be accounted for by polymorphism (CSAC 2005), which is likely to reduce the constraint estimates, since non-synonymous polymorphisms are less strongly affected by purifying selection than non-

synonymous divergence (Akashi 1995). This bias more strongly affects synonymous than non-synonymous constraint estimates, since the level of polymorphism at synonymous sites should be closer to the neutral level. As expected, I also found a strong negative correlation (r_P=-0.91, P=0.0003, df=9; see Table 3.5 and Figure 3.3) between constraint and generation time (Table 3.4), which is expected because of the known negative correlation between N_e and generation time (r_P=-0.88, P=0.00004, df=8; see Table 3.5) (Chao & Carr 1993; Ellegren 2009).

Surprisingly, synonymous constraint is not significantly correlated either with N_e (for $\log_{10}(N_e)$ r_P=0.11, P=0.76, df=8) (see Figure 3.2B) or with generation time (r_P=0.10, P=0.64, df=9) (see Figure 3.3B and Table 3.5).

3.4.3 Effect of divergence on selective constraint

For closely related bacteria (with a divergence of less than 2%), Rocha et al. (2006) found that the d_N/d_S ratio increased with the reciprocal of the divergence time, and showed that the strength of the relationship is positively influenced by N_e. The correlation was consistent with a simple model incorporating the time lag that is necessary for selection to remove slightly deleterious mutations at non-synonymous sites, i.e. standing polymorphism inflates d_N (see above). A similar correlation between d_N/d_S and divergence time was found in vertebrates (Wolf et al. 2009), but for much higher divergence values (between 0.02-3.04), where ancestral polymorphism is unlikely to play a role.

To test the effect of evolutionary distance on the constraint estimates, I analysed the correlation between constraint and the divergence measured from ARs. Non-synonymous constraint shows a relatively strong positive correlation with divergence (r_P=0.74, P<0.006, df=9) (see Figure 3.4A, Table 3.5), while for synonymous sites, I did not find a significant correlation overall (r_P=-0.40, P=0.22, df=9) (see Figure 3.4B, Table 3.5).

Table 3.4. Selective constraint estimates at non-synonymous and synonymous sites for eleven vertebrate species pairs.

Species pairs	Constraint		N_e	GT (years)
	Non-Synonymous	Synonymous		
human–chimpanzee	0.706 (0.698, 0.714)	0.210 (0.193, 0.226)	1.0×10^4 [2,8,9]	18 [1,3,5]
human–macaque	0.777 (0.772, 0.783)	0.210 (0.201, 0.219)	2.5×10^4 [2]	12 [1,3,5]
orangutan–gorilla	0.784 (0.778, 0.790)	0.239 (0.227, 0.252)	3.0×10^4 [2,8]	10 [1]
macaque–marmoset	0.805 (0.799, 0.811)	0.227 (0.218, 0.236)	4.2×10^4 [2]	4 [1,3,7]
mouse–rat	0.807 (0.802, 0.812)	0.117 (0.110, 0.123)	5.8×10^5 [9]	0.3 [1,3,7]
pig–cow	0.828 (0.821, 0.834)	0.096 (0.081, 0.110)	1.0×10^5 [2]	2.7 [3,6,7]
cow–alpaca	0.830 (0.818, 0.841)	0.188 (0.163, 0.211)	1.0×10^5 [2]	2 [3,7]
cow–dolphin	0.808 (0.801, 0.815)	0.201 (0.190, 0.211)	1.0×10^5 [2]	5 [3,7]
dog–cat	0.829 (0.819, 0.838)	0.200 (0.174, 0.222)	1.3×10^5 [2]	2.0 [3,7]
opossum–wallaby	0.858 (0.849, 0.867)	0.159 (0.139, 0.178)	-	1.5 [4,7]
chicken–turkey	0.871 (0.866, 0.876)	0.387 (0.377, 0.397)	6.0×10^5 [10]	2 [2]

Selective constraint was estimated at non-CpG-sites (see Materials and Methods). Effective population sizes (N_e) and generation time (GT) data are from the following sources: 1. Martin & Palumbi 1993, 2. Piganeau and Eyre-Walker 2009, 3. Bromham et al. 1996, 4. Kajin et al. 2008, 5. Elango et al. 2006, 6. Jezierski 1977, 7. Myers et al. 2006, 8. Yu et al. 2004, 9. Eyre-Walker et al. 2002, 10. Mank et al. 2010. In cases when values for N_e or GT were available for both species in the pair an average value was taken.

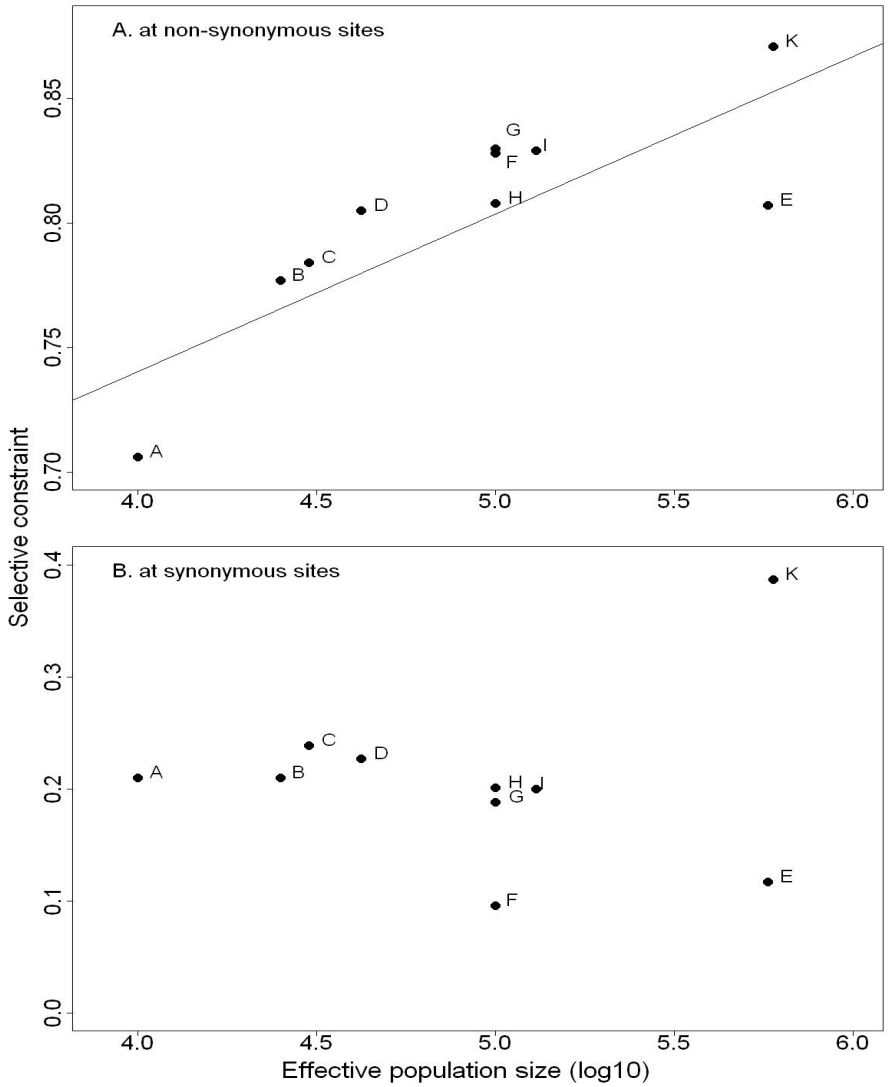

Figure 3.2. Relationship between selective constraint and the effective population size (N_e) in vertebrates. Non-synonymous constraint correlates positively with $\log_{10}(N_e)$ (r_P=0.83, P=0.014), but there is no significant correlation between synonymous constraint and $\log_{10}(N_e)$ (r_P=0.15, P=0.78). Correlation coefficients were calculated for the independent data points (B, F & G excluded). Selective constraint was estimated at non-CpG-prone sites. Abbreviations, constraint and N_e values are given in Table 1 and 2.

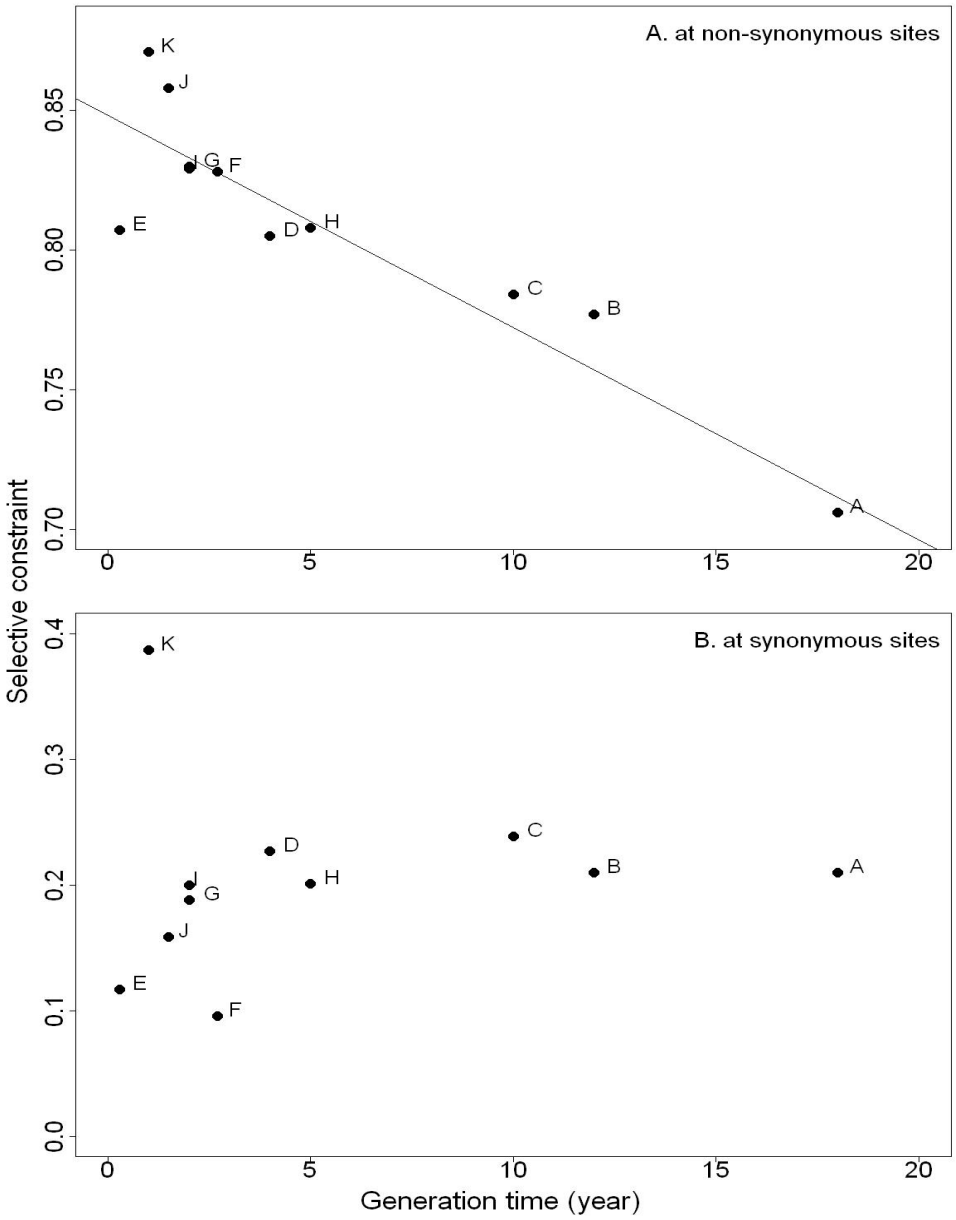

Figure 3.3. Relationship between selective constraint and generation time (*GT*) in vertebrates. There is a strong negative correlation between non-synonymous constraint and *GT* (rP=-0.91, P=0.0052), but no correlation between synonymous constraint and *GT*. (rP=0.015, P=0.78). Correlation coefficients were calculated for independent data points (i.e. B, F & G excluded). Abbreviations, *GT* and constraint values are given in Table 1 and 2.

Table 3.5. Correlations and partial correlations in the total and independent datasets.

Correlations	All			Without B,G &H			Without B, F & H			Without B, F & G		
	r	P	df	*r*	P	df	*r*	P	df	*r*	P	df
CN ~ log(N_e)	0.82	0.0025	8	0.82	0.0190	5	0.82	0.0202	5	0.83	0.01	5
CN ~ *GT*	-0.91	0.0003	9	-0.91	0.0075	6	-0.91	0.0079	6	-0.91	0.01	6
CN ~ *D*	0.74	0.0061	9	0.73	0.0324	6	0.72	0.0351	6	0.72	0.04	6
CS ~ log(N_e)	0.11	0.76	8	0.12	0.79	5	0.15	0.78	5	0.15	0.78	5
CS ~ *GT*	0.10	0.64	9	0.09	0.67	6	0.04	0.69	6	0.02	0.77	6
CS ~ *D*	-0.40	0.23	9	-0.41	0.31	6	-0.40	0.35	6	-0.38	0.36	6
N_e ~ *GT*	-0.88	0.0000	8	-0.88	0.0005	5	-0.87	0.0010	5	-0.89	0	5
D ~ *GT*	-0.79	0.0024	9	-0.76	0.0162	6	-0.76	0.0177	6	-0.75	0.02	6
Partial correlations												
CN ~ log(N_e), *D* fixed	0.67	0.0200	7	0.65	0.0793	4	0.69	0.0623	4	0.68	0.07	4
CN ~ log(N_e), *GT* fixed	0.10	0.4053	7	0.11	0.4232	4	0.12	0.4176	4	0.14	0.41	4
CN ~ *D*, log(N_e) fixed	0.40	0.1521	7	0.33	0.2660	4	0.38	0.2326	4	0.21	0.34	4
CN ~ *D*, *GT* fixed	0.09	0.4039	8	0.13	0.3968	5	0.11	0.4167	5	0.12	0.41	5
CN ~ *GT*, log(N_e) fixed	-0.69	0.0333	7	-0.70	0.1014	4	-0.71	0.0966	4	-0.66	0.11	4
CN ~ *GT*, D fixed	-0.78	0.0085	8	-0.80	0.0436	5	-0.80	0.0412	5	-0.8	0.04	5

CN : non-synonymous constraint, N_e: effective population size, *GT*: generation-time, *D*: divergence, All: all data points, B: human-macaque data, F: pig-cow data, G: cow-dolphin data, H:cow-alpaca data, *r*: Pearson correlation coefficient. *P*-values were calculated by bootstrapping the data 1000 times with shuffling.

The observed positive correlation between non-synonymous constraint and divergence (see Figure 3.4A), cannot be readily explained by the time lag model of Rocha et al. (2006) as most of the species pairs examined here are diverged to an extent where polymorphism is unlikely to affect substitution rate estimates. Instead, the observed relationship may to some extent be related to the insertion and removal of slightly deleterious mutations, the pace of which should depend on several factors like the generation time (Li et al. 1987; Ohta 1993), strength of selection ($N_e s$), the distribution of fitness effects of mutations (e.g. Boyko et al. 2008), their interactions (Kondrashov 1988), the mode of selection (Nielsen 2005) and mutational biases (McVean & Charlesworth 1999).

Indeed, when confounding factors, notably the generation time (*GT*), is controlled for, the correlation between non-synonymous constraint and evolutionary distance is reduced to almost zero (for fixed *GT*, r_P=0.09, P=0.40, df=8, see Table 3.5). This is because pairwise divergence correlates negatively with *GT* (r_P=-0.79, P=0.002, df=9; see Nikolaev et al. 2007), as first described by Ohta (1993) and led to the generation time hypothesis, which postulates that in taxa with higher *GTs* there are less germ-line cell divisions over a fixed time period, and so, less DNA replications which lead to less germ-line mutations (Li 1997). The previously observed negative correlations between the d_N/d_S ratio and both N_e and divergence are intimately linked to the effect of *GT*, as *GT* correlates negatively with N_e and the neutral substitution rate (see Table 3.5). In organisms with higher *GT*s, the mutation rate per year is lower (d_S is lower), while the proportion of effectively neutral mutations is higher (d_N is higher) due to their smaller N_e. These two factors lead to the correlation between the d_N/d_S ratio (non-synonymous constraint) and divergence, while the latter causing the correlation between the d_N/d_S ratio (non-synonymous constraint) and N_e.

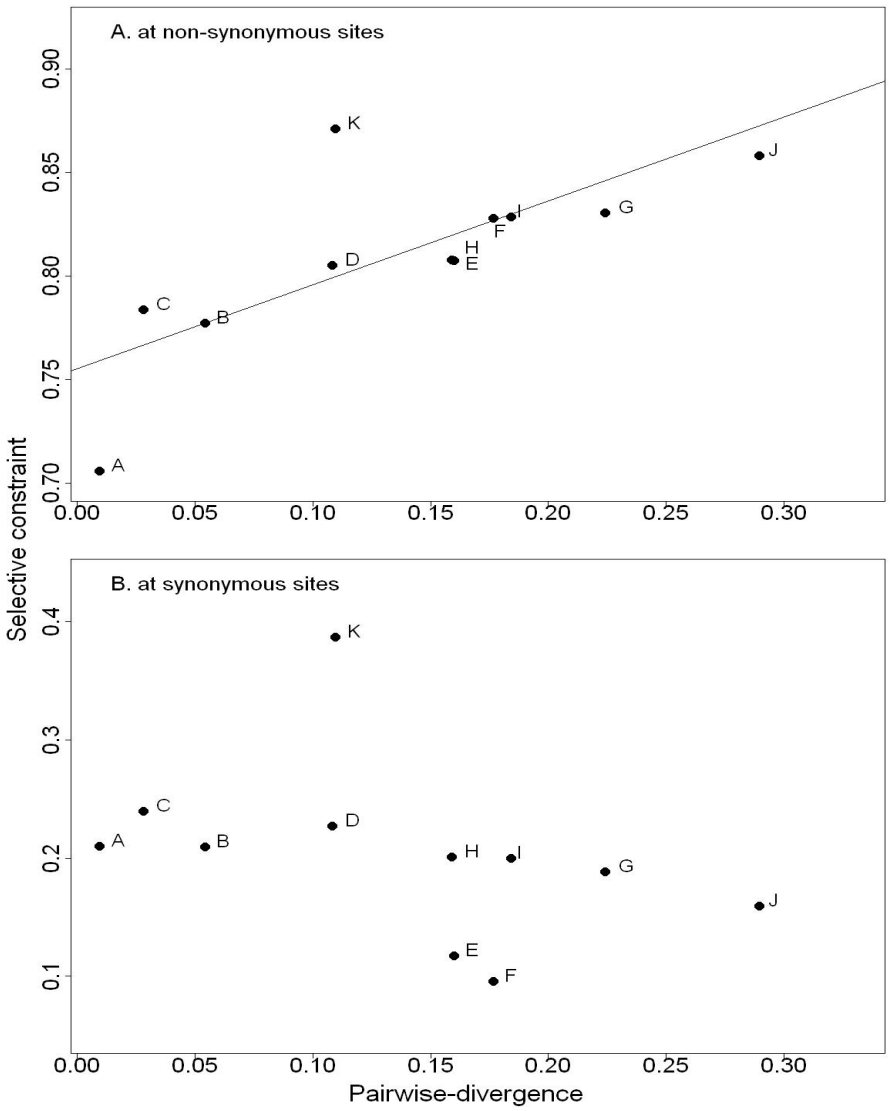

Figure 3.4. Relationship between selective constraint and the evolutionary distance, measured in ancestral repeats, in vertebrates. Non-synonymous constraint correlates positively with divergence (rP=0.72, P=0.041), but there is no significant correlation at synonymous sites (rP=-0.38, P=0.36). Correlation coefficients were calculated for independent data points (i.e. B, F & G excluded). Abbreviations, constraint and divergence values are given in Table 1 and 2.

It is important to note that testing for the effect of evolutionary distance on constraint estimates is not in itself tests for an effect of divergence time, a factor that may also affect constraint estimates.

3.4.4 Lack of correlations at synonymous sites

While the majority of mutations at non-synonymous sites are likely to be strongly deleterious because they affect the structure and/or function of the encoded protein, mutations at synonymous sites are more likely to have only mild effects on fitness (Chamary et al. 2006; Drummond & Wilke 2008). However, the true differences in the distribution of fitness effects between different functional categories are still largely unknown (Boyko et al. 2008). Nevertheless, it is reasonable to assume that N_e affects synonymous constraint, since the nearly neutral theory predicts a wider range for selection coefficients for mutations that behave effectively neutrally in species with larger N_e values, although mutational biases can obscure correlations between divergence and N_e when $N_e s<1$ (McVean & Charlesworth 1999). My results concerning generation time are compatible with what I obtained based on N_e, so I can exclude uncertainties in the magnitudes of estimates of N_e. (Estimates vary between 6,000 and 100,000 for human (Piganeau & Eyre-Walker 2009; Burgess & Yang 2008) and 160,000-580,000 for wild mice (Eyre-Walker et al. 2002; Halligan et al. 2010), and strongly depend on other potentially poorly known parameters, such as the mutation rate and population history.)

Although the synonymous constraint of 0.4 in birds may result from efficient selection acting on weakly deleterious mutations due to their large N_e, the question remains as to why rodents, with a similar N_e to birds, show consistently low constraint at synonymous sites (Gaffney & Keightley 2006; Chapter 2), even lower than that of marsupials, carnivores, ungulates and primates, all with undoubtedly lower N_e values (see Figure 3.2B and Figure 3.3B). Another question is why the groups of primates with the lowest N_e values have higher synonymous constraint

levels than the other vertebrate groups, apart from birds. The answer to these questions may be related to the following problems.

First, it is known that mammalian genomes are not at compositional equilibrium (see e.g. Duret & Arndt 2008). Although I have attempted to correct mutation rates for non-equilibrium processes, the fixed equilibrium GC content (p_e) which was assumed may not hold for the diverse set of species investigated here, nor may the assumption of constancy of p_e along chromosomes (for a discussion see Chapter 2). These assumptions may bias constraint estimates to a greater extent at synonymous sites, because their rate of evolution is considerably higher than that of non-synonymous sites, and their mean GC content is also higher by 2-12% (see Table 3.6). The only exception to this pattern is provided by birds, which have slightly lower synonymous GC content than at non-synonymous sites.

Table 3.6. Frequency of GC nucleotides in different sequence types.

Species pairs	GC content		
	Non-Synonymous	Synonymous	AR
human–chimpanzee	49.1%	54.9%	41.6%
human–macaque	49.3%	56.2%	39.4%
orangutan–gorilla	49.2%	55.5%	43.0%
macaque–marmoset	49.0%	54.0%	39.4%
mouse–rat	49.4%	56.2%	44.5%
pig–cow	49.5%	60.2%	40.2%
cow–alpaca	47.2%	48.9%	39.1%
cow–dolphin	49.7%	60.0%	41.1%
dog–cat	49.9%	62.3%	43.3%
opossum–wallaby	48.0%	52.5%	40.1%
chicken–turkey	47.1%	46.5%	48.0%

Second, it is known that biased gene conversion (*BGC*) can imitate the effect of selection (Nagylaki 1983) and is a major force in driving genome evolution of mammals (Duret & Arndt 2008). Indeed, it was found that as much as 20% of cases with significantly elevated d_N/d_S ratio in primates may be a result of *BGC* (Ratnakumar et al. 2010), but the extent to which *BGC* affects the estimates of synonymous and non-synonymous constraints in vertebrates is still largely unknown.

Third, there is evidence that shifts in population size may affect rates of substitutions when reverse mutations are taken into account. Such a process has been observed in cases of population expansion as well as contraction (Charlesworth & Eyre-Walker 2007), but effects on the signatures of selection are still poorly understood and depend on the population history and the fitness effects of mutations.

Fourth, the species included in my analysis have very different life-histories (e.g. generation time, body mass, metabolic rate), many of which are known to have an effect on sequence evolution (Martin & Palumbi 1993; Bromham et al. 1996; S. I Nikolaev et al. 2007), and may affect constraint estimates.

Last, while selection at non-synonymous sites largely acts to maintain or improve amino acid sequences, selection at synonymous sites is known to act on a diverse set of functions in mammals i.e. to preserve translational accuracy (Drummond & Wilke 2008) and translational efficiency (Waldman et al. 2010), maintain proper splicing through exon splice enhancer (ESE) and silencer (ESS) motifs (Parmley et al. 2006), and to maintain structural properties of mRNAs, which are necessary for their processing (Chamary & Hurst 2005). The significance of these processes may well be different between different species (e.g. the strength of selection acting on translational accuracy/efficiency; the number of different ESEs and ESSs motifs, their occurrence in sequences, and their turnover and selective pressure acting on them; or the number of nucleotides forming the stems in the stem-loop structures of RNAs) may contribute to the observed differences in synonymous constraints.

3.4.5 Concluding remarks

In general, the results presented here suggest that synonymous sites are under modest selective constraint. Averaging across species, about one-fifth of mutations at synonymous sites are eliminated by selection, which is likely to be underestimated, because sites evolving adaptively are unaccounted for.

The estimated constraint values and the differences in constraint (a value of 0.3) between species have the following implications for methods employing synonymous sites as neutral standards. First, constraint causes an underestimation of the mutation rate by a factor of 10-40%, depending on the species, and this can affect the results of fitting selection models based on an underlying phylogenetic relationship (e.g. Yang & Nielsen 2000; Pollard et al. 2010). Second, synonymous constraint affects the results of the dN/dS ratio (e.g. Kosiol et al. 2008; Wolf et al. 2009) causing it to be overestimated by 11-67%. For example the identification of genes that are under positive selection requires the dN/dS ratio to be over one. This can be caused both by positive selection at non-synonymous sites and by constraint at synonymous sites , as was shown by Chamary et al. (2006) and Wolf et al. (2009). This makes the McDonald-Kreitman test (McDonald & Kreitman 1991) for detecting adaptive evolution superior to the dN/dS ratio, because it is less affected by constraints on the assumed neutral standard. Third, the results of studies that aim to assess differences in mode and direction of selection between species (e.g. Bakewell et al. 2007; Kosiol et al. 2008) based on the dN/dS ratio is questionable, because the observed differences might be caused by differences in the level of constraint at synonymous sites between the compared species and not just by selection acting on non-synonymous sites.

4 Effect of non-equilibrium processes on inferences of selection, correcting the expectation under the neutral theory

4.1 Abstract

The extent to which selection drives the evolution of nucleotide sequences has been strongly debated since the formulation of Kimura's theory of neutral evolution. The theory has played an important role ever since, as it serves as the null hypothesis in tests designed to infer selection. These tests are, in general, based on the assumption that nucleotide sequences are at compositional equilibrium. Recent analyses suggest that, at least in mammals, sequences are evolving towards a lower equilibrium GC content. It is also well known that different sequence types, e.g. synonymous and non-synonymous sites, introns and untranslated regions have markedly different sequence compositions. These two facts raises the question as to what extent can non-equilibrium processes affect inferences of selection, i.e. whether the null hypothesis that the relative rate of evolution equals one in case of neutrality can be upheld. Here, I derive the deterministic equations describing the rate of evolution at sequences of different GC content which are not at compositional equilibrium, and discuss what effect could this have on selective constraint estimates. Biases caused by this process on the inferences on direction and strength of selection are also discussed.

4.2 Introduction

Inferences on the direction and strength of selection on stretches of genomic DNA, from single nucleotides to genes, chromosomes and whole genomes, are crucial to the understanding of the role of natural selection in maintaining the adapted state and driving the adaptation of species to their ever changing environments. Inferences

depend to a great extent on the assumption that there are regions within genomes which are non-functional, so that the observed substitution rates within these regions reflect the underlying mutational processes (i.e. mutations are either neutral or nearly neutral and, as a consequence, their fate largely depends on chance events; Kimura 1983; Ohta 1995). Comparing the within species polymorphism and/or between species divergence at the assumed neutral sites with the observed polymorphism and/or divergence at linked sites provides tests for neutral evolution (Miyata & Yasunaga 1980; McDonald & Kreitman 1991; Yang & Bielawski 2000).

For practical, and mainly historical reasons, synonymous sites, nucleotides at third codon positions where mutations do not affect the encoded amino acid sequence, have been widely used as local neutral standards (Kimura 1977; Ohta 1995) and most of our knowledge on selection acting at the level of individual genes is based on this assumption (Bakewell et al. 2007; Kosiol et al. 2008). As there is evidence that selection operates at synonymous sites (Chamary et al. 2006; Drummond & Wilke 2008), focus has shifted towards other sequence classes as putatively neutral sequences, at least for genome-wide studies. In rodents, Keightley & Gaffney (2003) found that, relative to introns, flanking intergenic regions, introns close to exons and even synonymous sites show signatures of negative selection. Similarly, all sequence types in hominids and murids (i.e. synonymous, non-synonymous, intron, 5'UTR, 3'UTR) appear to evolve under constraint when ancestral transposable elements (AR) serve as neutral standards (Gaffney & Keightley 2006; Pollard et al. 2010; Chapter 2).

The GC content of mammalian genomes shows regional variation, and it is well known that these genomes are made up of regions of relatively invariable GC content (see e.g. IHGSC 2001), frequently referred to as isochores (Bernardi 2000), the origin of which is still debated. Recent studies have indicated that mammalian nucleotide sequences are not at compositional equilibrium (Arndt et al. 2003; Khelifi et al. 2006; Duret & Arndt 2008), and they are evolving towards a lower equilibrium

GC content ($p_e < 0.5$), due to an excess of G or C to A or T (W: GC→AT) as opposed to A or T to G or C (S: AT→GC) substitutions. Indeed, it was shown that $W < S$ in different primate lineages (Duret & Arndt 2008; Ratnakumar et al. 2010). As a consequence, based on the directional mutation pressure theory of Sueoka (1962), a positive correlation between GC content and substitution rate is expected (Piganeau et al. 2002) and has been observed (Smith et al. 2002; Duret & Arndt 2008), although the observed relationship is not necessarily linear (IMGSC et al. 2002; Chapter 2).

As the GC content of different sequence types varies between 35 and 55%, being generally highest at synonymous sites and lowest for introns (e.g. (Hellmann 2003; Chapter 2), their total rate of evolution should also be different when they evolve towards a common equilibrium, even if they evolve in a largely neutral fashion. As a consequence, the substitution rates observed in a neutral sequence do not necessarily reflect the underlying neutral mutation rates. This may have serious implications for estimates of selection. For example, based on the relative rate test (Miyata & Yasunaga 1980), selection can be inferred in cases when two sequences are evolving completely neutrally, when the sequence chosen as the neutral reference does not have the same base composition as the target sequence for which selection is inferred. The direction and the strength of the inferred selection depends on the GC content of the two sequences at the start of the evolutionary process, and on the substitution rates, u and v. The facts that mammalian genomes are thought to be evolving towards a lower p_e and different sequence types have different GC content, make a correction for the neutral rates necessary if the direction and strength of selection is to be inferred.

Here, I derive equations describing the rate of evolution at sequences that are not at compositional equilibrium. Based on these equations and on the observed neutral substitution rates (W and S), I calculate the neutral mutation rates, u and v, for primates, and estimate p_e and the level of selective constraint for different sequence types. I demonstrate that the widely held belief that the null hypothesis for the relative rate test (i.e. that the relative rate of evolution for a sequence assumed to

evolve under selection is one in case of neutrality), may be wrong and should be adjusted according to the GC content of the neutral and the test sequence. Failing to do so may result in biases in the detection and quantification of the direction and strength of selection.

Based on the derived equations it is also possible to estimate what p_e would reach by any given sequence class, if the mutation rates and selection regimes remain constant over time, i.e. the equilibrium GC content of the various functional categories. I show that, contrary to the expectations Duret & Arndt (2008), different genic sequence types evolve towards a higher GC content at non-CpG sites; the only exception is observed at 5'UTRs.

The high constraint estimates at synonymous sites of hominids agree well with our previous results (Chapter 2 and 3), but also highlight the fact that, apart from non-synonymous sites, where constraint is generally high, level of constraint at other sequence types vary widely and may change rapidly.

4.3 Materials and Methods

Data. Three way alignments of human, chimpanzee and macaque, along with their reconstructed ancestral sequences were retained from the 6-way EPO primate alignments from Ensembl (Flicek et al. 2009) through the Perl API (Ensembl version 59). Ancestor alignments were generated by Ensembl based on Ortheus (Paten et al. 2008), which is an indel-aware progressive alignment method. Human annotations from Ensembl were mapped onto the alignments as described before (Chapter 2) and were used as a reference to mark the genes on both the chimp and macaque sequences. Genes with at least one putatively orthologous transcript in both the chimp and the macaque sequence were selected for the analysis. The criteria for valid transcripts were the following. A) transcripts should start in a start and end in a stop codon (transcripts which started/ended in the same codon as the reference human sequence were also retained); B) sequences were excluded when they experienced

frameshift mutations; and C) were not allowed to contain internal STOP codons, apart from those that code for selenocysteines. Given that the primate genome sequences are of good quality, and the species considered here are closely related, it is reasonable to assume that substitution rates are not biased and the effect of multiple hits is minimal, apart from the rate at hypermutable CpG dinucleotides, which may affect the rate estimates for macaque.

Data analysis.

Two site categories were distinguished: 1) sites which were not part of an ancestral CpG dinucleotide, referred to as non-CpG sites, and 2) those which, based on the reconstructed ancestral sequence, were part of a CpG site. Splitting the dataset in this way had the advantage that it was not necessary any more to exclude half of the total number of sites by restricting the analysis to non-CpG-prone sites only (sites not preceded by C or followed by G, see e.g. Meunier & Duret 2004) as I had done it before in Chapter 2 and 3. The multiple alignments made it possible to ascertain the direction of substitutions. For the first site category I distinguished between two pairwise substitution rates A↔T, G↔C and the two directional rates i.e. substitutions from G or C to A or T (*W*: GC→AT) and from A or T to G or C (*S*: AT→GC). At CpG sites only C→T or G→A rates were considered and these rates were used to include the contribution of CpG sites to the constraint estimates. Throughout this study I assumed that ancestral transposable elements, so called ancestral repeats (AR), evolve neutrally and the observed net substitution rates at these sites, *W* and *S*, were used to calculate the neutral mutation rates, termed *u* (GC→AT) and *v* (AT→GC) which are normalised by the frequency of GC and AT sites, respectively. *u* and *v* were derived from a deterministic model for the frequency of nucleotide changes, based on the work of Sueoka (1962). The model applies to the rate of substitution of mutations, including the effect of drift and selection, so the values of *u* and *v* can more generally be given as the product $uQ(-s, N_e)$ and $vQ(s, N_e)$, where *Q* is a function of selection and the effective population size (e.g.

Charlesworth & Eyre-Walker 2007). The assumptions of neutral evolution at *AR*s holds only if $|s| < 1/2N_e$.

Equations for the substitution rates are derived in the Theoretical Background section. Based on these equation, first the root of equation (7) was obtained by the bisection method (Teukolsky et al. 1988), which yields the sum of *u* and *v*. Then *v* and *u* were calculated from equation (6). These rates for ARs were used in turn to calculate the expected numbers of directional changes for any given sequence classes, i.e. for introns, synonymous and non-synonymous sites, 5' and 3' untranslated regions (5' or 3' UTRs). By substituting the neutral *u* and *v* and the GC content of the specific sequence types into equations (4) and (5), I obtained the expected rates *W* and *S*, i.e. the substitution rates that one would observe if the sequence types evolved neutrally, with a correction for the non-equilibrium process. Then, based on the four rates ($k_{i=1..4}$: A↔T, G↔C, AT→CG, GC→AT), the expected number of changes were obtained by multiplying the substitution rates by the number of sites where the specific substitution could have occurred following the equation of

$E = \sum_{i=1}^{4} k_i N_i$, where N_i is the number of A and T nucleotides for rates A↔T, AT→CG or G and C nucleotides for G↔C and GC→AT rates. Then by counting the number of observed changes (*O*), constraint was calculated as $C=1-O/E$.

Mutation rates are known to vary within and between chromosomes (Gaffney & Keightley 2006; Duret & Arndt 2008) and here I assume that this rate variation occurs at a scale of 1Mb. I also assume that within segments substitution rates do not vary with GC content. The rate and constraint calculations therefore were done by 1Mb segments, and 95% confidence intervals were obtained, by bootstrapping the dataset over the segments.

4.4 Results

In this study, I analysed the effect of non-equilibrium processes, i.e. when sequences with different GC content (p_0) evolve at different rates towards the equilibrium GC content (p_e). Here, I derive the equations that describe the rate of evolution in sequences which are not at compositional equilibrium, test the predictions of the model and apply it on primate genomic sequence data. Consequences of the predictions of the model and implications for different methods which are commonly used to infer strength of selection, notably the d_N/d_S ratio, are also discussed.

4.4.1 Theoretical background

It is well know that the sequence composition of different organisms varies in a wide range and that the change in the GC content of genomes, a very slow process, can be defined by the following first-order linear differential equation (Sueoka 1962), where p is the GC content of the sequence and u is the rate of change from G or C to A or T (GC→AT) per GC sites, and v is the rate of change from A or T to G or C (AT→GC) per AT sites (see also Materials and Methods).

$$\frac{dp}{dt} = v - (u+v)\,p$$

The solution to this equation (with initial value $p(0)=p_0$) is given as

$$p(t) = p_e + (p_0 - p_e)\,e^{-(u+v)t} \quad (1),$$

where p_e is the equilibrium GC content, the sequence would eventually reach if u and v remain constant over time and defined as $v/(u+v)$, and p_0 is the GC content at t_0 (e.g. Marais et al. 2004). The expected amount of change, $D(t)$, at any site in a nucleotide sequence over a time period t_1- t_0 can then be given as the solution to the following integral

$$D(t) = \int_{t_0}^{t_1} [\,p(t)u + (1-p(t))\,v\,]\,\mathrm{d}\,t \quad (2),$$

where $p(t)u$ is the rate of change for GC→AT and $(1-p(t))v$ is the rate of change for AT→GC per nucleotide at time *t*.

Substituting equation (1) into equation (2) leads to the following integral:

$$D(t)=\int_{t_0}^{t_1}[up_e+u(p_0-p_e)e^{-(u+v)t}]\,\mathrm{d}t+\int_{t_0}^{t_1}[v(1-p_e)-v(p_0-p_e)e^{-(u+v)}]\,\mathrm{d}t$$

$$=\frac{2uv}{u+v}(t_1-t_0)+\frac{v-u}{u+v}(p_0-p_e)(e^{-(u+v)t}-e^{-(u+v)t_0}) \quad (3)$$

For the frequency of GC→AT substitutions (*w*) per nucleotide:

$$W=\frac{uv}{u+v}(t_1-t_0)-\frac{u}{u+v}(p_0-p_e)(e^{-(u+v)t}-e^{-(u+v)t_0}) \quad (4)$$

For the frequency of AT→GC substitutions (*S*) per nucleotide:

$$S=\frac{uv}{u+v}(t_1-t_0)+\frac{v}{u+v}(p_0-p_e)(e^{-(u+v)t}-e^{-(u+v)t_0}) \quad (5)$$

Subtracting equation (4) from (5) and arbitrarily setting t_0=0 and t_1=1 (and also p_e= *v*/(*u*+*v*)), I obtain the following equation

$$\left(p_0-\frac{v}{u+v}\right)e^{-(u+v)}=S-W$$

Substituting *x*=*u*+*v* and *c*=*S*-*W,* I get

$$v=x(p_0-ce^x) \quad (6)$$

Using *x* and *c,* equation (5) can be rewritten as follows:

$$\frac{(x-v)v}{x}+\frac{v(p_0-v)(1-e^x)}{x^2e^x}-S=0 \quad (7)$$

now with the single unknown *x* (where *S* is a constant and *v* is a function of *x*).

The solution to this transcendental equation can be obtained numerically as described in Materials and Methods, which leads to the values of u and v, given observed values of p_0, S and W.

4.4.2 Model predictions

Focusing on the individual mutation (u, v) and substitution (W, S) rates, it follows from equation (3) that when $u = v$, (i.e. p_e=0.5), the substitution rate is constant at $W+S = 2uv / (u+v)$ for any p_0 over a fixed time interval. In all other cases, the observed substitution rate either linearly increases or decreases with p_0 depending on whether v less than or greater than u, respectively, if u and v are held fixed (see Piganeau et al. 2002 and Figure 4.1).

One consequence of the relationship between the observed substitution rates and p_0 is that one can infer selection for a completely neutral sequence when its rate of evolution is compared to that of another neutral reference. To illustrate this point let us consider a simple example. It is known that for the human genome $u > v$ for neutral rates, therefore the genome composition evolves towards a p_e which is estimated to be 0.37 (Duret & Arndt 2008). If we have a neutral reference with a given p_0 (e.g. p_0=0.42) and a fixed p_e such that $u > v$ (e.g. p_e=0.37), then $W + S$ is constant with a value greater than $2uv/(u+v)$. We then use the rate for a reference sequence (d_{ref}) to estimate the direction and strength of selection on another neutral sequence, called the target sequence, using the relative rate parameter (d_{tar}/d_{ref}), which is analogous to the widely used $\omega=d_N/d_S$. ratio. Depending on whether p_0 for the target sequence is less than or greater than that of the neutral reference, the inferred measure of selection is less than or greater than one, respectively (see Figure 4.1). The expected value of one is obtained only in two cases. First, when the neutral and the test sequence have exactly the same initial GC content (p_0). Second, in the special case when u=v, then the slope of the line is zero and for all p_0 values the relative rate is correctly given as one.

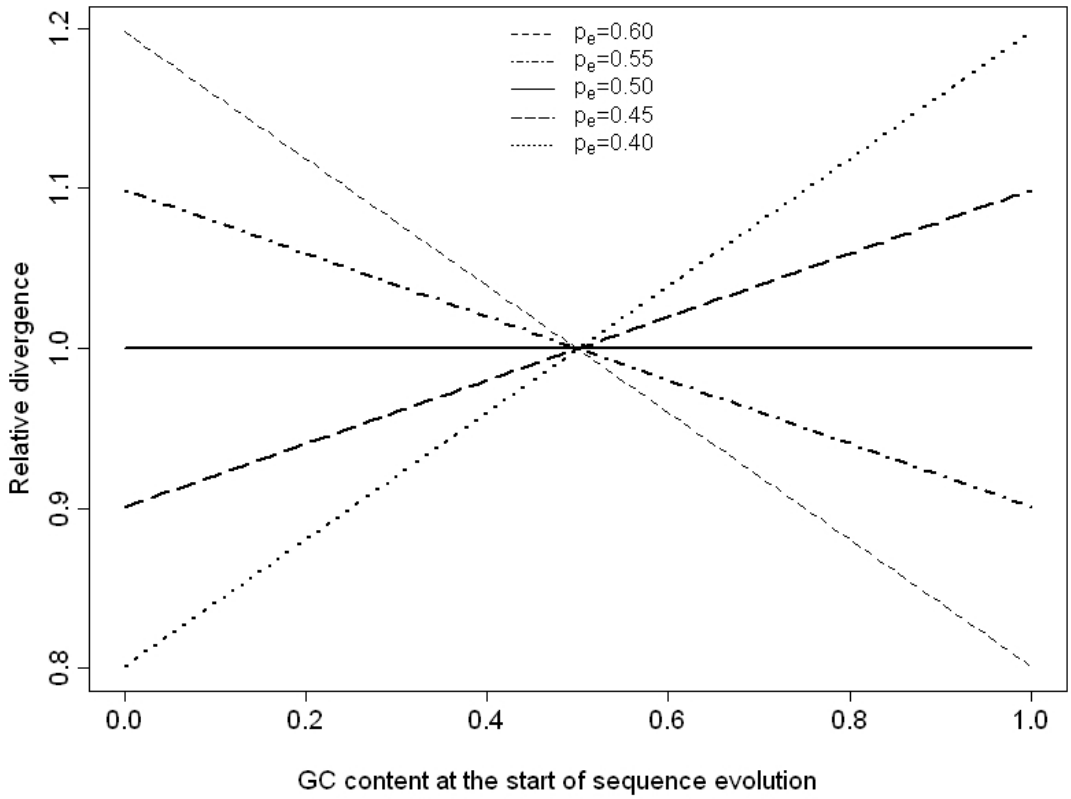

Figure 4.1. Effect of GC content on divergence. Divergence is the sum of the two directional substitution rates, *W*: GC→AT and *S*: AT→GC. The sum of the mutation rates, *u*: GC→AT and *v*: AT→GC and the time period is held fixed and the rate is given relative to the case when *u*=*v*, i.e. when the rate is $2uv/(u+v)$. The time interval is assumed to be 1 with t_0=0 and t_1=1.

By choosing a different sequence type for a neutral reference with a different p_0, then both the direction and the strength of the inferred selection may change (Figure 4.2).

In practice, this means that A) neutral sequences may appear to evolve under selection B) estimates of selection on functional sequences are biased when the neutral and selected sequence do not have the same ancestral base composition. The types of bias caused by non-equilibrium processes on inferences of selection by the relative rate test is shown in Table 4.1. The closer *u* and *v* to one another, the smaller

the gradient of the slope, and hence the bias caused (see dotted line on Figure 4.2 in comparison with the other lines). Whether the relative rate generates under- or over-estimates depends on the slope of the lines (positive or negative, depending on whether $u<v$ or $u>v$, i.e. $p_e>0.5$ or $p_e<0.5$, respectively) and on the p_0 values compared to the reference p_0.

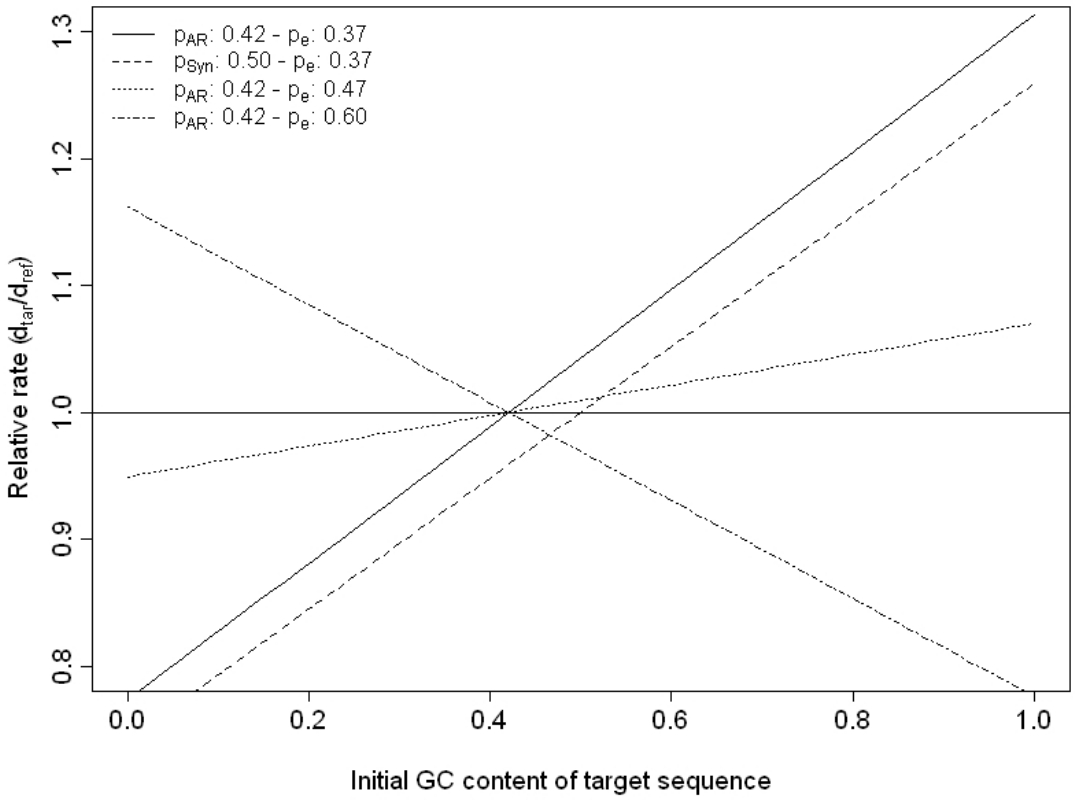

Figure 4.2. Effect of non-equilibrium processes on the relative rate parameter. Relative divergence values were calculated for a neutral target sequence (tar) relative to another neutral sequence called reference (ref) as a function of the GC content of the target. The equilibrium GC content (p_e) was set to three different values (0.37, 0.47, 0.60), keeping $u+v$ constant. GC content of the ref was fixed at two different values P_{AR} and P_{Syn}, where AR stands for ancestral repeat and Syn for synonymous sites.

Table 4.1. Biases on inferences of selection caused by non-equilibrium processes.

	p_0 tar < p_0 neut	p_0 tar > p_0 neut	p_0 tar = p_0 neut
$v<u$	underestimated	overestimated	unbiased
$v>u$	overestimated	underestimated	unbiased
$v=u$	unbiased	unbiased	unbiased

p_0 neut is the GC content of the assumed neutral sequence, p_0 tar is the GC content of the neutral target sequence for which is inferred selection.

4.4.3 Effect of non-equilibrium processes on d_N/d_S ratios

The previous assumption of constant GC content of the neutral standard, the case presented in Figure 4.2, is unrealistic. For example, in one of the most widely used method to infer seelction, the non-synonymous to synonymous rate ratio (d_N/d_S ratio), the direction and strength of selection on individual genes is measured by assuming that synonymous sites evolve neutrally (Yang & Bielawski 2000). The bias caused by non-equilibrium processes on the estimated selection in this case depends on the distribution of p_0 at synonymous and non-synonymous sites and the differences between u and v. Assuming that p_e is 0.37 in humans (Duret & Arndt 2008) and $u+v$ is around 0.01, the bias caused can be represented by a plane determined by two independent variables, synonymous and non-synonymous p_0, with the d_N/d_S ratio as the dependent variable, where d_N and d_S are calculated using equation (3). The obtained "neutral" d_N/d_S plane is shown in Figure 4.2, with human specific genes represented by dots. As it is suggested by the figure, with the given parameters, non-equilibrium processes have a strong effect on the estimated rate. The calculated neutral d_N/d_S ratios are in the range of 0.60 and 1.60. While these values are the most extreme, they clearly indicate that previous estimates could be biased, as the test statistics for the d_N/d_S ratio determine whether the obtained value is significantly

different from the assumed neutral value of one. Instead, in practice this value should be set based on the neutral expectation accounting for the non-equilibrium processes shown in Figure 4.2.

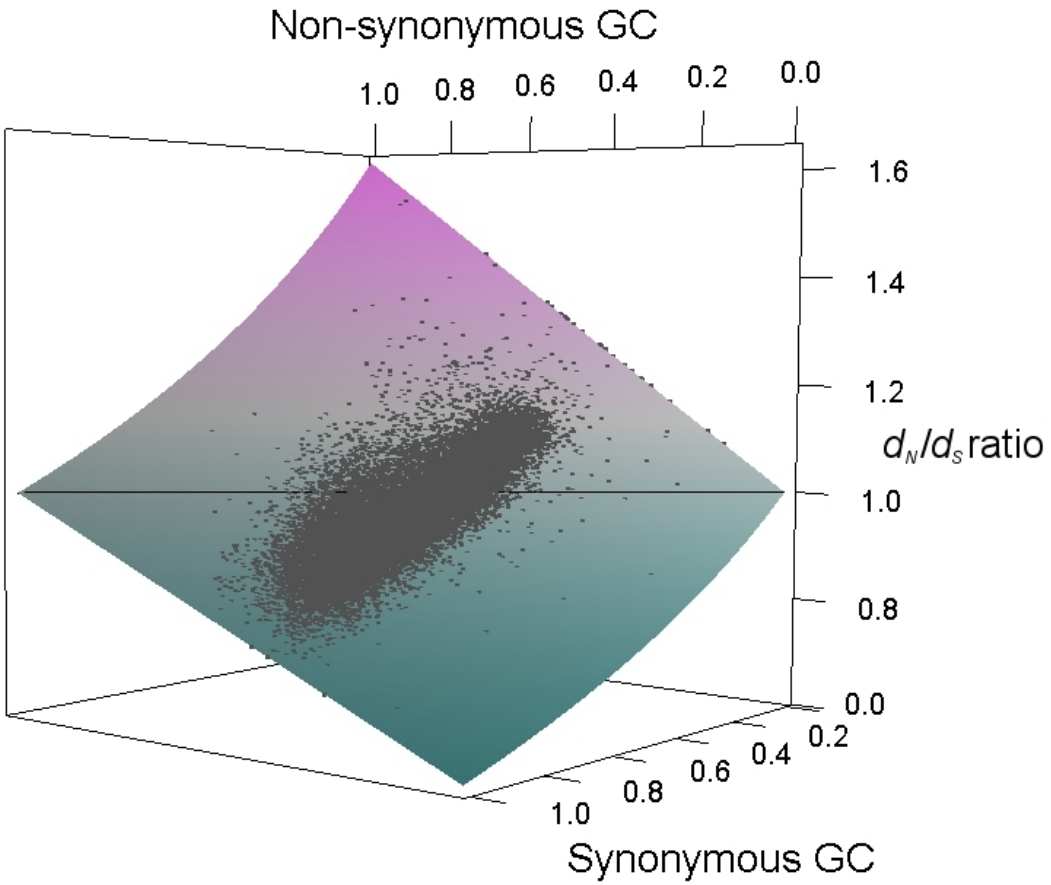

Figure 4.3. Bias in the estimates of direction and strength of selection caused by non-equilibrium processes on the d_N/d_S ratio. Assumed equilibrium GC content was 0.37 and the sum of GC→AT and AT→GC mutation rates were set to be 0.01. The d_N and d_S rates were calculated based on the equation derived in the Theoretical Background section. The d_N/d_S plane is then plotted as a function of GC content at synonymous and non-synonymous sites. Each dot corresponds to a single protein coding gene in the human genome. The d_N/d_S rates presented reflect the neutral expectation when both synonymous and non-synonymous sites evolve completely neutrally.

4.4.4 Accounting for non-equilibrium processes in primate constraint estimates

Recently, I have estimated selective constraint on different sequence types in hominids (see Chapter 1 and 2), and here I test the bias in these estimates when non-equilibrium processes are unaccounted for. Sequence data for human, chimp and macaque were collected from Ensembl (Flicek et al. 2009) and after filtering the data (see Materials and Methods) 10,776 genes were retained for analysis. A summary of the data is shown in Table 4.2.

The two pairwise (A↔T and G↔C) and two directional substitution rates (AT→GC and GC→AT) in ancestral repeats (AR) were estimated by comparing the nucleotides in the reconstructed ancestral to the nucleotides in the leaf sequence. Then, by substituting the directional rates and the GC content of the ARs into equation (7), I obtained the corresponding values of u and v, and p_e as $v/(u+v)$. Substitution rates at ARs are presented in Table 4.3, ancestral and equilibrium GC compositions are shown in Table 4.4. The estimated mean p_e is lowest for macaque and highest for human, with values between 0.42 and 0.47. These estimates are higher than anticipated from previous estimates on hominids (Arndt et al. 2003; Meunier & Duret 2004; Karro et al. 2008; Duret & Arndt 2008), which suggested that p_e is somewhere between 0.30 and 0.40. For example, the recent estimate of p_e of 0.37 for the human genome by Duret & Arndt (2008) is lower by a value of 10% than the estimate obtained here. More surprising is the fact that the equilibrium values of ARs are higher than their ancestral GC composition, which implies that they are evolving towards a higher GC content instead of a lower one.

To calculate average constraint on genic sequence types (i.e. on non-synonymous and synonymous sites, introns, 5' and 3' UTRs), I used the estimated neutral rates in ARs, to obtain corrected substitution rates for the specific sequence compositions of the different sequence types (see Materials and Methods). Corrected constraint estimates are shown in Table 4.5.

Table 4.2. Summary of data used in the analysis.

Lineages	Number of basepairs analysed (Mb)					
	AR	Intron	Non-synonymous	Synonymous	5'UTR	3'UTR
human	274.22	355.68	9.80	2.20	2.75	13.09
chimp	266.68	349.31	9.72	2.20	2.64	12.97
macaque	229.31	328.12	9.51	2.17	2.53	12.46

Data is given for 10,776 putatively homologous genes in primates.

Table 4.3. Substitution rate estimates for ancestral repeats.

A)

Lineages	A↔T	G↔C	*u* (GC→AT)	*v* (AT→GC)
human	0.00063 (0.00062, 0.00064)	0.00115 (0.00113, 0.00117)	0.00466 (0.00461, 0.00471)	0.00412 (0.00406, 0.00417)
chimp	0.00086 (0.00085, 0.00087)	0.00138 (0.00136, 0.00140)	0.00545 (0.00540, 0.00550)	0.00447 (0.00441, 0.00452)
macaque	0.00486 (0.00482, 0.00489)	0.00680 (0.00676, 0.00684)	0.03662 (0.03642, 0.03682)	0.02674 (0.02654, 0.02695)

B)

Lineages	d	CpG
human	0.00520 (0.00473, 0.00483)	0.04626 (0.04593, 0.04665)
chimp	0.00596 (0.00561, 0.00572)	0.05004 (0.04965, 0.05044)
macaque	0.03657 (0.03633, 0.03669)	0.18016 (0.17942, 0.18087)

Directional (A↔T and G↔C per AT and GC sites, respectively) and pairwise (GC→AT and AT→GC) substitution rates are shown in part A. *u* and *v* values were calculated using the deterministic equations to correct for non-equilibrium. Total divergences at non-CpG sites and at CpG sites are shown in part B. Unit of time for a species is given as the time separating the species from its most recent common ancestor. 95% confidence intervals are shown in parentheses.

Table 4.4. Frequency of G or C nucleotides and CpG dinucleotides in ancestral repeats.

Lineages	CpG-content	p_0	p_e
human	0.023	0.421	0.469
chimp	0.022	0.421	0.451
macaque	0.019	0.421	0.422

p_0 is the GC content of ancestral repeats (AR) at the start of evolution from the ancestral sequence. p_e is the equilibrium GC content.

It is possible to compare values of constraint estimates with and without correction for the non-equilibrium processes, but differences are so small that they do not exceed the value of 0.001 (data not shown). This is because average p_0 values (between 0.39-0.54) are relatively close to the neutral p_0 values (between 0.42-0.47) (see Table 4.4, 4.5) and the equilibrium values obtained are close to the value of 0.50 – when there is no bias in the estimates (see the dotted line representing the case for the human genome Figure 4.2). Taking the average of the human and chimp constraint values, I compared the results with my previous estimates (Chapter 2 and 3). Apart from 5'UTRs for which the new estimate are slightly lower, all the previous estimates are underestimates, but the differences are surprisingly small (between 0.005 and 0.06), given the different sets of genes, alignments, site types and method used. This verifies the previously described trends, i.e. the observed differences in level of constraints between different sequence types (see Chapter 2), but may affect the genome wide constraint estimates, as different types occur in the genome with different frequencies, and so their contribution to the amount of constrained sites is proportional to their representation in the genome.

Table 4.5. Selective constraint estimates for different sequence types in primates with corrections for non-equilibrium processes.

Lineages	Selective constraint				
	5'UTR	**Synonymous**	**Non-synonymous**	**Intron**	**3'UTR**
human	0.207 (0.206, 0.208)	0.195 (0.193, 0.195)	0.769 (0.768, 0.771)	0.049 (0.049, 0.049)	0.217 (0.216, 0.218)
chimp	0.120 (0.119, 0.122)	0.301 (0.300, 0.301)	0.781 (0.780, 0.781)	0.115 (0.114, 0.115)	0.285 (0.283, 0.285)
macaque	0.390 (0.390, 0.391)	0.290 (0.290, 0.291)	0.839 (0.839, 0.839)	0.155 (0.155, 0.155)	0.348 (0.348, 0.349)

95% confidence intervals are shown in parenthesis.

Comparing the lineage specific constraint estimates between human, chimp and macaque, I found that, while constraint differences at non-synonymous sites is relatively small (0.01 and 0.07 between human and chimp and human and macaque, respectively), average constraint estimates vary widely within the other categories. Differences in constraint between the three species frequently exceed 0.1 and in the case of 5'UTRs reaches 0.27 between chimp and macaque. These results are surprising, considering that I used a single set of orthologous genes and the time since the split of human and macaque is relatively small 25 Mya (Rogers et al. 2007).

As before, lowest constraint was found in introns, but while the results for human are consistent with previous estimates (Pollard et al. 2010; Chapter 2), the lineage specific constraint estimates are approximately twice and three times as high in the chimp and macaque, respectively. One possible explanation of the higher constraint values is that while the quality of annotation for the human genome is high, human annotations may not be correct for the chimpanzee and macaque genes, so that introns in the other two primate species may contain protein coding sequences. If this is the case, then there should be approximately four times as many hidden protein coding sites in macaque introns than suggested by the human annotations, which seems unlikely, especially that gene structure is assumed to be generally well conserved between different species (Roy et al. 2003). Deviations from the human annotation may, to some extent, cause the variation in constraint estimates at 5' and 3'UTRs, if the transcription start and end positions vary between species, but this could not explain the variation at synonymous sites.

Differences in effective population sizes and as a consequence in the proportion of mutations evolving effectively neutrally (Ohta 1995, see also Chapter 3) may to some extent contribute. As the long term effective population sizes (N_e) were suggested to be around 10,000 in human, 20,000 in chimp and 40,000 in macaque (Yu et al. 2004; Piganeau & Eyre-Walker 2009), therefore the expectations based on N_e are higher constraint values moving from human towards macaque. This is indeed

observed at non-synonymous sites, introns and 3'UTRs, but not at the other two sequence categories. These expectations depend, in a great extent, on the uncertainties in N_e estimates and may also be affected by standing polymorphisms, which can cause lower constraint estimates, at least in human and chimp (for a discussion see Chapter 2).

The lineage specific results also suggest that synonymous sites evolve under constraint, and the level of constraint is even higher in chimp and macaque than I previously estimated in pairwise comparisons (Chapter 2 and 3). The observed variations in the estimates of constraints are unlikely to be caused by mutational biases (Majewski 2003; Green et al. 2003) or processes associated with recombination i.e. higher mutation rates (Lercher & Hurst 2002; Arnheim & Calabrese 2009), increased efficiency of section (Felsenstein 1974) e.g. through the Hill-Robertson interference (Hill & Robertson 1966) or biased gene conversion (Meunier & Duret 2004; Duret & Arndt 2008) as these should have about the same effect on our neutral standards. Instead it may suggest that selective pressure on synonymous sites and on other categories may change rapidly.

Beside calculating selective constraint, the derived equations, make it possible to calculated p_e for the different functional categories (Table 4.6) which represent the equilibrium composition that functional sequences would eventually reach if mutational forces and selection regimes remained the same. In general, differences between p_0 and p_e values are between zero and 0.1, but while 5'UTRs evolve towards lower p_e values, the other sequence types are either in equilibrium or evolve towards higher GC contents. This observation along with the fact that most of the sequence types retain high GC contents and in general evolve towards higher equilibrium than ARs may suggest a role for selection in specifying the composition of sequences. While GC-biased gene conversion (BGC) is known to affect the evolution of GC content (Duret & Arndt 2008), this is unlikely to cause the observed differences between p_e values at different sequence types, as BGC should affect the neutral and functional sequences similarly. Neither can mutational biases cause the observed

trends, as these shown regional variation, and all of the sequence types examined come from coding sequences, which are likely to be equally affected by transcription coupled mutation and repair mechanisms (Green et al. 2003; Majewski 2003).

Table 4.6. GC content and equilibrium GC content for different sequence types.

Lineages	5'UTR		Synonymous		Non-synonymous		Intron		3'UTR	
	p_0	p_e	p_0	p_e	p_0	p_e	p_0	p_e	p_0	p_e
human	0.535	0.464	0.498	0.479	0.461	0.496	0.386	0.456	0.406	0.475
chimp	0.533	0.430	0.498	0.500	0.461	0.465	0.385	0.445	0.406	0.460
macaque	0.538	0.473	0.503	0.533	0.461	0.464	0.387	0.425	0.407	0.449

p_0 is the GC content of the sequence type at non-CpG sites, p_e is the equilibrium GC content that the sequence would reach if mutational processes and intensity of selection remain constant.

4.5 Discussion

Although the estimated p_e values for ARs (between 0.42 and 0.47) may seem to be in contradiction with the previous estimates, which suggest much lower p_e values for primates (Arndt et al. 2003; Duret & Arndt 2008), in practice there is a crucial difference between the different results. The equations derived here are based on the assumption that mutations at nucleotide sites occur independently from one another. This assumption is certainly violated, at least in vertebrates, since it is well known that mutation rate at a single nucleotide sites may depend on the immediate neighbour nucleotides (this effect is called context dependency). The most well known example is the so called CpG hypermutability (Bird 1980), which concerns dinucleotides where nucleotide C is immediate followed by nucleotide G. At these sites substitution rate was estimated to be between 8 and 18 fold higher than at non-CpG sites in mammals (e.g. Arndt et al. 2003; Lunter & Hein 2004) due to the CpG methylation deamination process (Bird 1980). The effect of CpG hypermutability is also observed on the data analysed here with 5-9 fold higher rates at CpG sites than at non-CpG sites (see Table 4.3).

While the method described here can be used to infer p_e at non-CpG sites, which are assumed to be unaffected by CpG hypermutability, the inferred values do not hold for the total genome. Although the equilibrium composition of 0.47 holds for 98% of the transcribed genome, the remaining 2% of CpG sites inflate the GC→AT substitution rate only, and contribute with another 50% of GC→AT substitutions to the total. Calculating p_e based on the non-CpG S (0.0040) and W (0.0052) substitution rates, estimated by Duret & Arndt (2008), leads to an equilibrium of 0.44, which is now much closer to our estimate than their genome-wide estimate of 0.37. This suggest the followings:

First, at non-CpG sites compositional equilibrium is relatively close to 0.50 and around 98% of the genome evolves towards a higher GC content which is most likely

to be driven by BGC (Meunier & Duret 2004; Duret & Arndt 2008), and in some cases (at functional sites) by selection.

Second, CpG hypermutability is the main factor causing mammalian genomes to evolve towards lower equilibrium, but this process crucially depends on the birth and death of CpG dinculeotides, which in turn depends on p_0 (Jabbari & Bernardi 2004), mutational bias at CpG sites (Arndt et al. 2003; Lunter & Hein 2004), the effect of BGC (Meunier & Duret 2004; Duret & Arndt 2008), level of methylation (Ehrlich et al. 1982; Law & Jacobsen 2010) and selection at functional sites (Ohlsson 2002).

Third, the decrease and increase in the frequency of G and C nucleotides at CpG and non-CpG sites, respectively, will determine the isochore structure of mammalian genomes.

Last, the p_e values at non-CpG sites in primates suggest that the effect of non-equilibrium processes is much smaller in this fraction of the genome than I previously estimated based on an equilibrium of 0.37. This makes the d_N/d_S planes much flatter than it is shown in Figure 4.2, and so the biases at these sites are very small (see Table 4.7). It is important to note, that while the bias on estimates of selection at non-CpG sites is small it may be substantial when all sites are considered. As the model described here does not include context dependency the methods used to infer selection based on the d_N/d_S ratio should either be extended to account for this effect, e.g. similarly to the approaches of Hwang & Green (2004) or Duret & Arndt (2008), or should be restricted to sites which are not prone to the effect of context dependent mutations.

Table 4.7. Biases on d_N/d_S estimates of primates caused by non-equilibrium processes.

Lineages	range for d_N/d_S	median (d_N/d_S<1)	median (d_N/d_S>1)
human	0.91 – 1.12	0.98	1.01
chimp	0.86 – 1.20	0.97	1.02
macaque	0.79 – 1.32	0.95	1.03

d_N/d_S ratios for assumed neutrally evolving sequences were calculated based on the neutral rates of u and v by applying equation (3) with the initial GC content at synonymous and non-synonymous sites of protein coding genes annotated in the human genome.

The lineage specific constraint results presented here agree well with my previous estimates, but also highlight the fact that there is substantial variation in constraint within different sequence types. This may be important, not just because constraint at synonymous sites causes the d_N/d_S ratio to be overestimated, but because the results show that the u/v ratio is closer to one at synonymous sites, which may be the consequence of selection. A comparison of p_e values at ARs and the equilibrium towards which synonymous sites evolve show that the equilibrium is higher at synonymous sites by values of 0.01 for human, 0.05 for chimp and 0.11 for macaque (see Table 4.4 and 4.6). Although for human and chimp the difference is small, for macaque it is substantial, and the observed p_e which exceeds 0.5 is the consequence of a rate difference such that $v>u$. This is a peculiar feature of macaque, which is only observed at synonymous sites, while for all the other sequence types in macaque $u<v$. Something similar is observed at synonymous sites in chimp, where the results suggest that $u=v$. To what extent these deviations from the generally observed $u<v$ relation can be attributable to the effect of selection remains to be a question, but it was suggested that nucleotide C may be under selection for stable mRNA structure

(Chamary & Hurst 2005), and selection for exon splice enhancer sequences is known to affect the base composition at synonymous sites (Parmley et al. 2006). Deviations from the commonly observed relationship of $u>v$ at synonymous sites in primates raises the question as to how reliable conclusions can be made on the dynamics of GC content evolution based on these sites (Romiguier et al. 2010) and to what extent would this reflect for the real dynamics of GC content evolution in mammalian and other genomes.

While the results presented here suggest that the three primate species analysed here are not much affected by non-equilibrium processes (at non-CpG sites), the effect on other species can be substantial. In a recent analysis Romiguier et al. (2010) reported that GC content evolve differently in 33 mammalian genomes and the trends generally observed in placental mammals are different from that of hominids and murids. Invertebrate species can also be affected. For example, it is well known that the genome of Drosophila melanogaster is not at compositional equilibrium (Kern & Begun 2005), and there is a marked difference between observed and expected GC content at synonymous sites, 0.65 and 0.32, respectively, based on polymorphism data (Singh et al. 2007). If the substitution rates at synonymous sites reflect the real mutational pressure towards a much lower equilibrium, and GC content is different between synonymous and non-synonymous sites than the biases caused by non-equilibrium processes on the relative rate test may be even higher than presented here in Figure 4.2.

4.5.1 Concluding remarks

The d_N/d_S ratio is frequently used to infer selection at the genic level or below, with the aim of identifying genes or regions of genes which may have been the target of adaptive evolution (Nielsen et al. 2005; Bakewell et al. 2007; Kosiol et al. 2008). As I have shown above, these results should be treated with caution. First, the results presented here and in Chapter 2 and 3 show that synonymous sites are under constraint (between 20 and 30% of mutations are eliminated by selection at these sites in mammals), and as a result the d_N/d_S ratio overestimates the strength of

selection. Second, selection may modify the u/v ratio at synonymous sites in a way which may no longer reflect the underlying neutral mutational pressures. Last, in general non-equilibrium processes may bias estimates of selection, and the d_N/d_S ratio may be prone to this, as GC content varies widely within and between synonymous sites. The level of this bias depends on the mutational bias, and in primates does not have a strong effect on estimates, as u/v is close to one, when sites affected by CpG hypermutability are excluded.

Nevertheless the model highlights that the bias caused by non equilibrium processes can be substantial and also points to the importance of accounting for CpG hypermutability on estimates of selection. It can be readily applied for species with no known context dependent mutation effects, like Drosophila, or in cases when context dependency is accounted for (e.g. by excluding those sites which are prone to such processes).

5 Discussions and conclusion

5.1 Summary of the results

The evolution of nucleotide sequences is driven by recurrent mutations and the effect of random genetic drift and selection acting upon them (e.g. Kimura 1983). Understanding the properties of new mutations, for example the rate with which they arise (e.g. Kondrashov 2003; Lynch 2010b), the biases associated with them (e.g. Sueoka 1962; Bird 1980), as well as the fitness effects of mutations (e.g. Eyre-Walker & Keightley 2007; Boyko et al. 2008) and their interactions, are therefore crucial in the study of molecular evolution.

In Chapter 2, I estimated the level of selective constraint, i.e. the fraction of selectively eliminated mutations, on different protein coding, and non-coding sequence types in two mammalian taxa, hominids and murids. Comparing the human and chimpanzee, and the mouse and rat genomes, I used the observed substitution patterns in ancestral repeats (ARs) to estimate the expected number of changes in sequence types for comparison with the corresponding observed numbers. As expected, most deleterious mutations occurred at non-synonymous sites, where mutations change the amino acid sequence and by doing so may render the protein non-functional, or even potentially harmful. Selective constraint on the other sequence types is lower and varies more widely.

The results suggest that more functional sites are associated with alternatively spliced genes than with single transcript genes, probably to specify the proper splicing pattern of the mRNA, or to maintain the timing, tissue specificity or level of expression of the different isoforms.

Contrary to a previous study that suggested that upstream intergenic regions of genes may have lost their functional significance in hominids (Keightley et al. 2005), I found that the level of constraint on upstream intergenic regions (within 5kb) of

genes is approximately the same between hominids and murids, although constraint values are significantly lower at 5'UTRs. The difference between the two studies is at least partly caused by the previous assumption that introns are selectively neutral, although this is unlikely to be the case since these sites appear to be weakly constrained relative to ARs.

One of the most interesting results is that in hominids, more than one fifth of the mutations at synonymous sites are eliminated by selection, but only a tenth in murids. This observation somewhat contradicts the expectation from the nearly neutral theory (Ohta 1995), which predicts a larger fraction of borderline mutations to be eliminated by selection in species with higher N_e values, i.e. higher constraint is expected on murid sequence types, as N_e in rodents is approximately 60 times higher than N_e in hominids (e.g. Piganeau & Eyre-Walker 2009; Halligan et al. 2010). The observed reverse trend is not confined to synonymous sites, but is also observed in the downstream flanking intergenic regions and to some extent in introns. The causes of the observed differences in level of constraint between the two taxa are unknown, but these may either be method related biases or could even be the consequence of a reorganisation of functional sites.

Mean constraint results on sequence types are very close to the estimates which were recently obtained by statistical phylogenetic tests by Pollard et al. (2010). Lower bounds for the fraction of constrained nucleotides in the two genomes were calculated based on the mean constraint values for different sequence classes. The estimates of 5.5% are in good agreement with the estimate of 5% obtained in the pilot phase of a large scale study which aims to catalogue empirically defined functional elements in the human genome (Birney et al. 2007).

With estimates for the genome wide constraint and mutation rate, I estimated the rate of deleterious mutations per generation to be 4.4 for the diploid genome of hominids and 0.6 for murids. Based on a multiplicative model for the fitness effect of mutations (Kimura & Maruyama 1966) the mutation load on the primate populations

would be incredibly high, 99%. In a recent study on mutational effects in humans, Lynch (2010) argued that, even if the fitness effect of these mutations are small, the total effect of recurrent mutations may reduce the fitness by 1% per generation if selection is relaxed, which would lead to noticeable reduction in fitness in human populations within a few centuries, although this effect not necessarily permanent, as the present N_e for humans is very large and can keep the frequencies of deleterious alleles at a low level.

The observed moderate constraint level provides evidence for selection at hominid synonymous sites, and contradicts to the widely held assumption that synonymous sites can serve as neutral standard for inferences on direction and strength of selection.

The two-fold difference in synonymous constraint between hominids and murids suggest that constraint may vary widely, and also that the non-synonymous to synonymous substitution rate ratio (d_N/d_S) may not only lead to overestimates, but the bias caused by selection may vary from species to species. This may make the results on interspecies comparisons of the prevalence of adaptive evolution based on d_N/d_S ratios (e.g. Bakewell et al. 2007; Kosiol et al. 2008) questionable. To study the extent of this variation I analysed, in Chapter 3, eleven closely related species pairs and found that while non-synonymous constraint varies between 0.71 – 0.87, the range for synonymous constraint is almost twice as wide, 0.10 – 0.39. As a consequence differences in the inferred direction and strength of selection between species are likely to be more strongly affected by deviations from the neutral expectations at synonymous sites, than being caused by differences at non-synonymous sites.

Previous studies have shown that the d_N/d_S ratio correlates with the divergence time (measured by d_S) (e.g. Wolf et al. 2009) and N_e (e.g. Ellegren 2009); therefore in Chapter 3, I tested the effect of the divergence rate at ARs (d_{AR}), and N_e on synonymous and non-synonymous constraint estimates. At non-synonymous sites, I found that according to the expectations, constraint positively correlated with both

d_{AR} and N_e. These correlations are frequently handled independently from one another (e.g. Wolf et al. 2009), i.e. without considering an interaction between divergence and N_e, although it is well known that there is a negative correlation between generation time (*GT*) and N_e (Chao & Carr 1993). The generation time hypothesis (Li et al. 1987; Ohta 1993) postulates that there is a negative correlation between divergence and *GT*. Interestingly, for all data points, the partial correlation between non-synonymous constraint and divergence is greatly reduced and becomes non-significant when N_e is fixed, while it remaines high and significant when d_{AR} is fixed. Therefore, it is likely that the main factor causing the correlation is N_e (and hence probably the effect of selection), while the effect of divergence is to a great extent caused by the correlations between N_e and divergence because both correlate with *GT*. It should be noted however, that partial correlations (although remain high between non-synonymous constraint and N_e, when d_{AR} is fixed) are non significant when the data sets are reduced to phylogenetically independent data points.

Surprisingly, no correlation was found between constraint at synonymous sites with either d_{AR} or N_e, although a positive correlation was expected at least with N_e based on the expectation of the nearly neutral theory (Ohta 1995).

Another interesting result from the vertebrate analysis is that the highest constraint levels at both synonymous and non-synonymous sites were obtained for birds (chicken-turkey pair). Although it has been suggested that synonymous sites in mammals are under purifying selection, this is the first study which shows, that mutations at these sites are selected against in birds, and that constraint is almost twice as high in chicken and turkey as in mammals.

One problem with the analyses presented in Chapter 2 and 3 is the lack of information on the direction of changes, which is an inherent problem with pairwise sequence comparisons. If for example there is an excess of G or C to A or T (*u*: GC→AT) as opposed to A or T to G or C (*v*: AT→GC) substitutions, then sequences with higher actual GC content (p_0) experience more changes than those with lower

p_0, as a consequence of the mutational bias. This can lead to inferences of selection even in cases when rates of evolution are compared in two neutrally evolving sequences but with different p_0.

Mammalian sequences are to evolve towards lower equilibrium GC contents (p_e), i.e. there is a substitutional bias in mammals, such that $u > v$ (e.g. Khelifi et al. 2006; Duret & Arndt 2008), so non-equilibrium processes may cause a bias on estimates of selection.

In Chapter 4, I derived the deterministic equations specifying the value of nucleotide changes in sequences which are not at compositional equilibrium. These equations provide a way of obtaining corrected values for the expected number of nucleotide changes for a given u, v and p_0, by taking into account the non-equilibrium processes. Using multiple sequence alignments for primates (human, chimp and macaque) and the reconstructed ancestral sequences I estimated constraint levels on the different genic sequence types and compared the corrected values with the results in Chapter 2 and 3. Although I expected a sizeable error in the previous estimates, they are very close to the new constraint values (relative error is less than 0.01), in spite of the differences in the number and set of orthologous genes and site types.

The low errors seem to be the consequence of a u very close to v, which leads to a much higher p_e of 0.47 than the genome-wide estimate of 0.37 of Duret & Arndt (2008). There are at least two factors behind this difference. First, my estimates were obtained for intronic ARs, which are subject to processes associated with transcription coupled mutation and repair (Majewski 2003; Green et al. 2003). Second, CpG sites, which make up around 2% of ARs were excluded from the analysis. As all of the mutations at CpG sites are GC→AT, so excluding these sites reduces u, which leads to much higher p_e values than at all sites. As introns and intronic ARs make up more than 90% of the transcribed portion of the genome and their p_0 is lower than 0.47 at non-CpG sites, this means that they evolve towards a higher GC content which is most likely caused by GC biased gene conversion (BGC)

(Meunier & Duret 2004; Duret & Arndt 2008). The results also highlights the importance of CpG hypermutability in driving the genome towards a lower equilibrium GC content, and also, that the dynamics of GC content evolution probably depends on BGC and the birth and death of CpG dinucleotides.

Based on the deterministic equations, I estimated the base composition that the different sequence classes are evolving towards, as a consequence of mutational biases and selection. I found that, while 3'UTRs and non-synonymous sites are evolving towards a lower GC content, there is a reverse trend for 5'UTRs. GC content evolution is most varied at synonymous sites, with human evolving to a higher, macaque to a lower GC composition, while synonymous sites in chimp seems to be at compositional equilibrium.

5.2 Future directions

The mutation load in mammals, estimated in Chapter 2 depends to a large extent on the real number of functional sites in the genome and on the estimate of mutation rate. While the method employed here to estimate the fraction of functional sites within mammalian genomes seems to provide reasonable estimates, it is also restricted mainly by two factors. First, the assumption that transposable elements (TEs) evolve neutrally is unrealistic. Their roles for example in gene expression regulation (Pereira et al. 2009), exonisation and intronisation (Sorek et al. 2004; Lin et al. 2008) are well documented, therefore it is likely that the neutral rates obtained here are underestimated. Additionally, mutational biases may depend to some extent on the effect of selection, while the per generation mutation rate largely depends on the assumptions on divergence and generation times. Ultimately these uncertainties can be excluded when direct estimates of mutation rates in mammals (Kondrashov 2003; Lynch 2010a) become available at a genome-wide manner. Second, constraint estimates based on comparisons of observed and expected changes cannot account for the fraction of functional sites which are adaptively evolving. This can be overcome by making a catalogue of all functional elements within genomes (Birney

et al. 2007) and testing their mode of evolution on deep coverage multiple sequence alignments of mammalian genomes (e.g. Pollard et al. 2010).

Constraint estimates at different sequence types, presented in Chapter 2, 3 and 4, show large scale variation within and between species, but what the functions or processes selection is operating on is largely unknown. At synonymous sites, for example, it is known that selection can operate on codon usage bias for translational efficiency or accuracy (Rocha 2006; Drummond & Wilke 2008), on splicing signals like exon splice enhancers and silencers (e.g. Parmley et al. 2007) and on mRNA structure (Chamary & Hurst 2005). Another possibly important, but so far unexplored process, which can cause constraint at synonymous sites is regulation. Recent functional annotations (e.g. Flicek et al. 2009) suggest that transcription factor and other regulatory binding sites overlap with protein coding sequences, and a recent study indicated that binding sites are under functional constraint (Gaffney et al. 2008). The contribution of these processes to estimates of constraint is unknown but may well be explain the observed differences in synonymous constraint in vertebrates.

Finally, in Chapter 4, I presented a method which provides a correction to the null hypothesis of neutral sequence evolution for non-equilibrium processes. Although this method, based on deterministic equations, cannot be easily extended to include the effect of context dependent mutational processes, the effect of any such process can be tested in species with no context dependency, or in cases when context dependency is accounted for (e.g. by excluding CpG sites). While in hominids my results at non-CpG sites show that this effect is small, it may be larger in different mammals as their GC content, at least at synonymous sites seems to evolve differently (Romiguier et al. 2010). For chicken, Duret & Arndt (2008) predicted that pe may be low, around 0.39, based on recombination rates and on substitution rates assumed to be similar to that of humans. This prediction can now be tested on the available chicken, turkey and duck sequences. It is also known that the Drosophila melanogaster genome is not at equilibrium and is evolving from a p_0 of 0.65 towards

a p_e of 0.32 (Singh et al. 2007), according to polymorphism data at synonymous sites. For such parameters, the bias is likely to exceed the values which were given in Chapter 4 for humans, assuming pe to be 0.37, as the average p_0 in Drosophila is around 0.35 (Halligan & Keightley 2006). The results presented in the chapter also highlight the importance of CpG hypermutability on the evolution of GC content in mammalian genomes, and suggest that codon based substitution models (e.g. Yang & Bielawski 2000) may lead to biases in the estimates of the direction and strength of selection if context dependency and non-equilibrium processes are not accounted for.

6 References

Akashi, H., 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics*, 139(2), pp.1067-1076.

Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062), pp.1149-1152.

Arndt, P.F., Petrov, D.A. & Hwa, T., 2003. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Molecular Biology and Evolution*, 20(11), pp.1887-1896.

Arndt, P.F. & Hwa, T., 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, 21(10), pp.2322-2328.

Arnheim, N. & Calabrese, P., 2009. Understanding what determines the frequency and pattern of human germline mutations. *Nat Rev Genet*, 10(7), pp.478-488.

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*, 104(18), p.7489.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. *Gene*, 241(1), pp.3-17.

Bird, A.P., 1986. CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067), pp.209-213.

Bird, A.P., 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7), pp.1499-1504.

Birney, E. et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799-816.

Blencowe, B.J., 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends in Biochemical Sciences*, 25(3), pp.106-110.

Boyko, A.R. et al., 2008. Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLoS Genetics*, 4(5), p.e1000083.

Bray, N. & Pachter, L., 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Research*, 14(4), pp.693-699.

Bromham, L., Rambaut, A. & Harvey, P.H., 1996. Determinants of rate variation in mammalian DNA sequence evolution. *Journal of molecular evolution*, 43(6), pp.610–621.

Burgess, R. & Yang, Z., 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular biology and evolution*, 25(9), p.1979.

Bustamante, C.D., Nielsen, R. & Hartl, D.L., 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Molecular Biology and Evolution*, 19(1), pp.110-117.

Cartwright, R.A., 2009. Problems and solutions for estimating indel rates and length distributions. *Molecular Biology and Evolution*, 26(2), pp.473-480.

Castillo-Davis, C.I. et al., 2002. Selection for short introns in highly expressed genes. *Nature Genetics*, 31(4), pp.415-418.

Chamary, J. & Hurst, L.D., 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biology*, 6(9), p.R75.

Chamary, J., Parmley, J.L. & Hurst, L.D., 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews. Genetics*, 7(2), pp.98-108.

Chao, L. & Carr, D.E., 1993. The Molecular Clock and the Relationship between Population Size and Generation Time. *Evolution*, 47(2), pp.688-690.

Charlesworth, B., 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3), pp.195–205.

Charlesworth, B. & Charlesworth, D., 1998. Some evolutionary consequences of deleterious mutations. *Genetica*, 102, pp.3–19.

Charlesworth, J. & Eyre-Walker, A., 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proceedings of the National Academy of Sciences*, 104(43), pp.16992 -16997.

Chatterjee, S. & Pal, J.K., 2009. Role of 5'- and 3'-untranslated regions of mRNAs in human diseases. *Biology of the Cell*, 101(5), pp.251-262.

Chiaromonte, F. et al., 2003. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harbor Symposia on Quantitative Biology*, 68, pp.245-254.

Cleveland, W.S. & Devlin, S.J., 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403), pp.596-610.

Costantini, M., Cammarano, R. & Bernardi, G., 2009. The evolution of isochore patterns in vertebrate genomes. *BMC genomics*, 10(1), p.146.

Crow, J.F., 2000. The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics*, 1(1), pp.40–47.

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

Darwin, C., 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.*, John Murray, London.

Dermitzakis, E.T. & Clark, A.G., 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Molecular Biology and Evolution*, 19(7), p.1114.

Dorus, S. et al., 2004. Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell*, 119(7), pp.1027-1040.

Drummond, D.A. & Wilke, C.O., 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2), pp.341-352.

Duret, L. & Arndt, P.F., 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics*, 4(5).

Ehrlich, M. et al., 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, 10(8), pp.2709 -2721.

Elango, N., Thomas, J.W. & Yi, S.V., 2006. Variable molecular clocks in hominoids. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5), pp.1370-1375.

Ellegren, H., 2009. A selection model of molecular evolution incorporating the effective population size. *Evolution; International Journal of Organic Evolution*, 63(2), pp.301-305.

Eory, L., Halligan, D.L. & Keightley, P.D., 2010. Distributions of Selectively Constrained Sites and Deleterious Mutation Rates in the Hominid and Murid Genomes. *Molecular Biology and Evolution*, 27(1), pp.177-192.

Eyre-Walker, A., 2006. The genomic rate of adaptive evolution. *Trends in Ecology & Evolution*, 21(10), pp.569-575.

Eyre-Walker, A. & Keightley, P.D., 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution*, 26(9), pp.2097-2108.

Eyre-Walker, A. & Keightley, P.D., 1999. High genomic deleterious mutation rates in hominids. *Nature*, 397(6717), pp.344-347.

Eyre-Walker, A. & Keightley, P.D., 2007. The distribution of fitness effects of new mutations. *Nature Reviews. Genetics*, 8(8), pp.610-618.

Eyre-Walker, A. et al., 2002. Quantifying the slightly deleterious mutation model of molecular evolution. *Molecular biology and evolution*, 19(12), p.2142.

Felsenstein, J., 1974. THE EVOLUTIONARY ADVANTAGE OF RECOMBINATION. *Genetics*, 78(2), pp.737-756.

Fisher, R.A., 1930. *The Genetical Theory of Natural Selection* 1st ed., Oxford University Press, Oxford.

Fisher, R.A. & Yates, F., 1948. *Statisticsal tables for biological, agricultural and medical research.* 3rd ed., London: Oliver & Boyd.

Flicek, P. et al., 2009. Ensembl's 10th year. *Nucleic Acids Research*.

Frith, M.C., 2006. Evolutionary turnover of mammalian transcription start sites. *Genome Research*, 16(6), pp.713-722.

Gaffney, D.J., Blekhman, R. & Majewski, J., 2008. Selective Constraints in Experimentally Defined Primate Regulatory Regions. *PLoS Genet*, 4(8), p.e1000157.

Gaffney, D.J. & Keightley, P.D., 2008. Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. *BMC Evolutionary Biology*, 8, p.265.

Gaffney, D.J. & Keightley, P.D., 2006. Genomic Selective Constraints in Murid Noncoding DNA. *PLoS Genetics*, 2(11), p.e204.

Gaffney, D.J. & Keightley, P.D., 2005. The scale of mutational variation in the murid genome. *Genome research*, 15(8), p.1086.

Green, P. et al., 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics*, 33(4), pp.514-517.

Haag-Liautard, C. et al., 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature*, 445(7123), pp.82-85.

Haddrill, P.R. et al., 2005. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. *Genome Biology*, 6(8), p.R67.

Hahn, M.W., Demuth, J.P. & Han, S., 2007. Accelerated Rate of Gene Gain and Loss in Primates. *Genetics*, 177(3), pp.1941-1949.

Haldane, J.B.S., 1957. The cost of natural selection. *Journal of Genetics*, 55(3), pp.511-524.

Halligan, D.L. et al., 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. *Genome Research*, 14(2), pp.273-279.

Halligan, D.L. & Keightley, P.D., 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Research*, 16(7), pp.875-884.

Halligan, D.L. et al., 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genetics*, 6(1), p.e1000825.

Han, M.V. et al., 2009. Adaptive evolution of young gene duplicates in mammals. *Genome Research*, 19(5), pp.859-867.

Havlioglu, N. et al., 2007. An Intronic Signal for Alternative Splicing in the Human Genome S. Maas, ed. *PLoS ONE*, 2(11), p.e1246.

Hellmann, I., 2003. Selection on Human Genes as Revealed by Comparisons to Chimpanzee cDNA. *Genome Research*, 13(5), pp.831-837.

Hill, W.G. & Robertson, A., 1966. The effect of linkage on limits to artificial selection. *Genetical Research*, 8(3), pp.269-294.

Hobolth, A. et al., 2007. Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLoS Genetics*, 3(2), p.e7.

Hubbard, T. et al., 2007. Ensembl 2007. *Nucleic Acids Research*, 35, pp.D610-617.

Hubbard, T. et al., 2009. Ensembl 2009. *Nucleic Acids Research*, 37, pp.D690-D697.

Hudson, R.R., Kreitman, M. & Aguade, M., 1987. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics*, 116(1), pp.153-159.

Hughes, T.A., 2006. Regulation of gene expression by alternative untranslated regions. *Trends in Genetics*, 22(3), pp.119-122.

Hurst, L.D., 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*, 18(9), pp.486-487.

Hwang, D.G. & Green, P., 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39), p.13994.

IHGSC, 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860-921.

IHMC, 2005. A haplotype map of the human genome. *Nature*, 437(7063), pp.1299-1320.

IMGSC et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp.520–562.

Jabbari, K. & Bernardi, G., 2004. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*, 333, pp.143-149.

Jensen-Seaman, M.I. et al., 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research*, 14(4), pp.528-538.

Johnson, J.M. et al., 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 302(5653), pp.2141-2144.

Josse, J., Kaiser, A.D. & Kornberg, A., 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *The Journal of Biological Chemistry*, 236, pp.864-875.

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), p.2740.

Karro, J.E. et al., 2008. Exponential Decay of GC Content Detected by Strand-Symmetric Substitution Rates Influences the Evolution of Isochore Structure. *Molecular Biology and Evolution*, 25(2), pp.362 -374.

Keightley, P.D. & Eyre-Walker, A., 2000. Deleterious mutations and the evolution of sex. *Science (New York, N.Y.)*, 290(5490), pp.331-333.

Keightley, P.D. & Eyre-Walker, A., 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177(4), pp.2251-2261.

Keightley, P.D. & Gaffney, D.J., 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23), p.13402.

Keightley, P.D., Lercher, M.J. & Eyre-Walker, A., 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biology*, 3(2), p.e42.

Kern, A.D. & Begun, D.J., 2005. Patterns of Polymorphism and Divergence from Noncoding Sequences of Drosophila melanogaster and D. simulans: Evidence for Nonequilibrium Processes. *Molecular Biology and Evolution*, 22(1), pp.51 -62.

Khelifi, A. et al., 2006. GC Content Evolution of the Human and Mouse Genomes: Insights from the Study of Processed Pseudogenes in Regions of Different Recombination Rates. *Journal of Molecular Evolution*, 62(6), pp.745-752.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2), pp.111-120.

Kimura, M., 1968a. Evolutionary Rate at the Molecular Level. *Nature*, 217(5129), pp.624-626.

Kimura, M., 1968b. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research*, 11(3), pp.247-269.

Kimura, M., 1962. On the Probability of Fixation of Mutant Genes in a Population. *Genetics*, 47(6), pp.713-719.

Kimura, M., 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608), pp.275-276.

Kimura, M., 1983. *The neutral theory of molecular evolution*, Cambridge Univ Pr.

Kimura, M. & Crow, J.F., 1964. The Number of Alleles That Can Be Maintained in a Finite Population. *Genetics*, 49(4), pp.725-738.

Kimura, M. & Maruyama, T., 1966. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54(6), p.1337.

King, J.L. & Jukes, T.H., 1969. Non-Darwinian evolution. *Science (New York, N.Y.)*, 164(881), pp.788-798.

Kondrashov, A.S., 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human mutation*, 21(1), pp.12–27.

Kondrashov, A.S., 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *Journal of Theoretical Biology*, 175(4), pp.583-594.

Kondrashov, A.S., 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336(6198), pp.435-440.

Kondrashov, A.S. & Crow, J.F., 1993. A molecular approach to estimating the human deleterious mutation rate. *Human Mutation*, 2(3), pp.229-234.

Kosiol, C. et al., 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genetics*, 4(8), p.e1000144.

Law, J.A. & Jacobsen, S.E., 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews. Genetics*, 11(3), pp.204-220.

Lercher, M.J. & Hurst, L.D., 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics: TIG*, 18(7), pp.337-340.

Lercher, M.J. et al., 2003. A unification of mosaic structures in the human genome. *Human Molecular Genetics*, 12(19), pp.2411-2415.

Lewontin, R.C. & Hubby, J.L., 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics*, 54(2), pp.595-609.

Li, W-H., 1997. *Molecular evolution*, Sinauer Associates.

Li, W-H., Gojobori, T. & Nei, M., 1981. Pseudogenes as a paradigm of neutral evolution. *Nature*, 292(5820), pp.237-239.

Li, W-H., Tanimura, M. & Sharp, P.M., 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *Journal of Molecular Evolution*, 25(4), pp.330-342.

Lin, L. et al., 2008. Diverse splicing patterns of exonized Alu elements in human tissues. *PLoS Genetics*, 4(10), p.e1000225.

Lowe, C.B., Bejerano, G. & Haussler, D., 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences*, 104(19), p.8005.

Lunter, G. & Hein, J., 2004. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20, pp.i216-i223.

Lunter, G., Ponting, C.P. & Hein, J., 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol*, 2(1), p.e5.

Lynch, M., 2010a. Evolution of the mutation rate. *Trends in Genetics*, 26(8), pp.345-352.

Lynch, M., 2010b. Inaugural Article: Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences*, 107(3), pp.961-968.

Lynch, M., 2006. The Origins of Eukaryotic Gene Structure. *Molecular Biology and Evolution*, 23(2), pp.450 -468.

Lynch, M., 2007. *The origins of genome architecture*, Sunderland (MA): Sinauer Associates.

Lynch, M., Scofield, D.G. & Hong, X., 2005. The evolution of transcription-initiation sites. *Molecular Biology and Evolution*, 22(4), pp.1137-1146.

Lytle, J.R., Yario, T.A. & Steitz, J.A., 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proceedings of the National Academy of Sciences*, 104(23), p.9667.

Majewski, J., 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. *The American Journal of Human Genetics*, 73(3), pp.688–692.

Makalowski, W. & Boguski, M.S., 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 95(16), pp.9407-9412.

Mank, J.E. et al., 2010. Effective population size and the faster-X effect: empirical results and their interpretation. *Evolution*, 64(3), pp.663–674.

Marais, G., Charlesworth, B. & Wright, S.I., 2004. Recombination and base composition: the case of the highly self-fertilizing plant Arabidopsis thaliana. *Genome Biology*, 5(7), pp.R45-R45.

Margoliash, E. & Smith, E.L., 1965. *Evolving genes and proteins (Bryson, Vernon; Vogel, Henry J.; eds.)* V. Bryson & H. Vogel, eds., New York: Academic Press. Available at: http://dx.doi.org/10.1021/ed043pA544.3 [Accessed December 8, 2010].

Martin, A.P. & Palumbi, S.R., 1993. Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, 90(9), pp.4087-4091.

McDonald, J.H. & Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328), pp.652-654.

McVean, G.A. & Charlesworth, B., 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genetics Research*, 74(02), pp.145–158.

Meunier, J. & Duret, L., 2004. Recombination drives the evolution of GC-content in the human genome. *Molecular Biology and Evolution*, 21(6), p.984.

Miller, W. et al., 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Research*, 17(12), pp.1797-1808.

Miyata, T. & Yasunaga, T., 1980. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution*, 16(1), pp.23-36.

Moses, A.M. et al., 2006. Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila. *PLoS Computational Biology*, 2(10), p.e130.

Muller, H.J., 1950. Our load of mutations. *American Journal of Human Genetics*, 2(2), pp.111-176.

Nachman, M.W. & Crowell, S.L., 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1), p.297.

Nagylaki, T., 1983. Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences of the United States of America*, 80(20), p.6278.

Nielsen, R., 2005. Molecular signatures of natural selection. *Annual Review of Genetics*, 39, pp.197-218.

Nielsen, R. et al., 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, 3(6), p.e170.

Nikolaev, S.I. et al., 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proceedings of the National Academy of Sciences*, 104(51), p.20443.

Ohlsson, R., 2002. New Twists on the Epigenetics of CpG Islands. *Genome Research*, 12(4), pp.525-526.

Ohta & Gillespie, 1996. Development of Neutral and Nearly Neutral Theories. *Theoretical Population Biology*, 49(2), pp.128-142.

Ohta, T., 1993. An examination of the generation-time effect on molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 90(22), pp.10676-10680.

Ohta, T., 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, 246(5428), pp.96-98.

Ohta, T., 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution*, 40(1), pp.56-63.

Parmley, J.L., Chamary, J. & Hurst, L.D., 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular Biology and Evolution*, 23(2), pp.301-309.

Parmley, J.L. et al., 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol*, 5, p.e14.

Paten, B. et al., 2008. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Research*, 18(11), pp.1829-1843.

Patterson, N. et al., 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097), pp.1103-1108.

Pereira, V., Begun, D. & Eyre-Walke, A., 2009. The effect of transposable element insertions on gene expression evolution in rodents. *PLoS One*, 4(2).

Piganeau, G. & Eyre-Walker, A., 2009. Evidence for variation in the effective population size of animal mitochondrial DNA. *PloS One*, 4(2), p.e4396.

Piganeau, G. et al., 2002. Expected Relationship Between the Silent Substitution Rate and the GC Content: Implications for the Evolution of Isochores. *Journal of Molecular Evolution*, 54(1), pp.129-133.

Pollard, K.S. et al., 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1), pp.110-121.

Ponting, C.P., 2008. The functional repertoires of metazoan genomes. *Nature Reviews Genetics*, 9(9), pp.689–698.

Ramensky, V.E. et al., 2008. Positive selection in alternatively spliced exons of human genes. *American Journal of Human Genetics*, 83(1), pp.94-98.

Ratnakumar, A. et al., 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552), pp.2571 -2580.

Reuter, M. et al., 2008. A test of the null model for 5' UTR evolution based on GC content. *Molecular Biology and Evolution*, 25(5), pp.801-804.

Rocha, E.P., 2006. The quest for the universals of protein evolution. *Trends in Genetics*, 22(8), pp.412–416.

Rocha, E.P. et al., 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of theoretical biology*, 239(2), pp.226–235.

Rogers, J. et al., 2007. Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science*, 316(5822), pp.222-234.

Romiguier, J. et al., 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Research*, 20(8), pp.1001 -1009.

Roy, S.W., Fedorov, A. & Gilbert, W., 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proceedings of the National Academy of Sciences*, 100(12), p.7158.

Schneider, A. et al., 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biology and Evolution*, 2009(0), p.114.

Schwartz, S. et al., 2003. Human–mouse alignments with BLASTZ. *Genome Research*, 13(1), p.103.

Shabalina, S.A. & Spiridonov, N.A., 2004. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology*, 5(4), p.105.

Siepel, A. et al., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8), pp.1034-1050.

Siepel, A. & Haussler, D., 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, 21(3), pp.468-488.

Singh, N.D. et al., 2007. Patterns of Mutation and Selection at Synonymous Sites in Drosophila. *Molecular Biology and Evolution*, 24(12), pp.2687 -2697.

Smith, N.G.C., Webster, M.T. & Ellegren, H., 2002. Deterministic mutation rate variation in the human genome. *Genome Research*, 12(9), pp.1350-1356.

Sorek, R., 2003. Intronic Sequences Flanking Alternatively Spliced Exons Are Conserved Between Human and Mouse. *Genome Research*, 13(7), pp.1631-1637.

Sorek, R. et al., 2004. Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Molecular Cell*, 14(2), pp.221-231.

Su, A.I. et al., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16), p.6062.

Subramanian, S. & Kumar, S., 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, 168(1), pp.373-381.

Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America*, 48(4), pp.582-592.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), pp.585-595.

Takai, D. & Jones, P.A., 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6), pp.3740 -3745.

Takano-Shimizu, T., 1999. Local recombination and mutation effects on molecular evolution in Drosophila. *Genetics*, 153(3), pp.1285-1296.

Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P., 1988. *Numerical Recipes in C: The art of scientific computing*, Cambridge university press.

Urrutia, A.O., 2003. The Signature of Selection Mediated by Expression on Human Genes. *Genome Research*, 13(10), pp.2260-2264.

Vignaud, P. et al., 2002. Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature*, 418(6894), pp.152–155.

Vinogradov, A.E. & Anatskaya, O.V., 2007. Organismal complexity, cell differentiation and gene expression: human over mouse. *Nucleic Acids Research*, 35(19), p.6350.

Waldman, Y.Y. et al., 2010. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Research*, 38(9), pp.2964 -2974.

Wallace, B. & Dobzhansky, T., 1962. Experimental proof of balanced genetic loads in Drosophila. *Genetics*, 47, pp.1027-1042.

Wang, E.T. et al., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), pp.470-476.

Webster, M.T., Smith, N.G.C. & Ellegren, H., 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. *Molecular Biology and Evolution*, 20(2), pp.278-286.

Wetterbom, A. et al., 2006. Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. *Journal of Molecular Evolution*, 63(5), pp.682-690.

Wolf, J.B.W. et al., 2009. Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. *Genome Biology and Evolution*, 1, pp.308-319.

Wright, S., 1931. Evolution in Mendelian Populations. *Genetics*, 16(2), pp.97-159.

Yang & Bielawski, 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution (Personal Edition)*, 15(12), pp.496-503.

Yang, J., Su, A.I. & Li, W.H., 2005. Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. *Molecular Biology and Evolution*, 22(10), p.2113.

Yang, Z. & Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, 17(1), p.32.

Yu, N. et al., 2004. Nucleotide diversity in gorillas. *Genetics*, 166(3), p.1375.

Zuckerkandl, E. & Pauling, L., 1965. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins*, pp.97–166.