Mutation and Selective Constraint in the Murid Genome

Daniel J. Gaffney

PhD

University of Edinburgh

May 24, 2006

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own, except where explicitly stated otherwise in the text.

This work has not been submitted for any other degree or professional qualification.

Daniel Gaffney, May 24, 2006

To Mum and Dad, for everything.

Acknowledgements

This thesis would never have existed without the help of many people. I am indebted to all of the following.

Peter Keightley for supervision, encouragement and helpful comments on all the work in this thesis. Brian Charlesworth for additional supervision, advice and comments on thesis chapters. Ian White for statistical help. Dan Halligan for many discussions and comments on thesis chapters. Penny Haddrill, Jonathan Coe and Asher Cutter for helpful comments on thesis chapters.

To all my friends in Edinburgh Dan, Penny, Tim, Eileen, Jon, Stu K, Stu McGregor, Ali, Grainne, Allan, Carl, Alex, Jordi, Assumpta, Dave, George, Jules, Vincente, Richard, Drennan, James and Katie. Most of all thanks to Holly.

To all my friends in Ireland and elsewhere John, Ursula, Sarah, Fogo, Trish, Glen, Carmel, Aidan, Brendan, Paddy, Hannah, Mandy, Louise, Fi and Ciara.

Finally, I thank my parents, Joy and John. This thesis would not have been possible without you.

Publications

The following published papers have arisen from this thesis.

- Keightley, P. D. and Gaffney, D. J. 2003 Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents *Proc. Nat. Acad. Sci. USA* **100** 13402-13406
- Gaffney, D. J. and Keightley, P. D. 2004 Unexpected conserved non-coding DNA blocks in mammals *Trends Genet.* **20** 332-337
- Gaffney, D. J. and Keightley, P. D. 2005 The scale of mutational variation in the murid genome *Genome Res.* **15** 1086-1094

Abstract

A large proportion of the genome of many higher eukaryotes consists of apparently functionless noncoding DNA, the significance of which is a long-standing puzzle in biology. The aim of this work was to quantify the extent to which both mutation and natural selection have influenced molecular evolution in murid noncoding sequence. In particular, the magnitude of and variation in selective constraint within murid noncoding DNA was investigated. Selective constraint is defined as the proportion of all mutations occurring at a locus or site which are strongly deleterious and therefore removed by natural selection. The approach adopted to estimate selective constraint relies on the assumption that we can quantify the past strength of purifying selection in a DNA sequence by comparison with nearby regions which are assumed to be evolving neutrally. To this end, work in this thesis deals both with mutational variation and bias (Chapters 2 and 3) as well as with selective constraint (Chapters 4 and 5) in noncoding DNA.

Chapter 2 is concerned with the differential effects of context-dependent mutation (namely, CpG hypermutability) at fourfold synonymous and noncoding sites. Using simulations it was shown that a common method of assigning ancestral CpG status often introduces a substantial level of bias into the estimation of nucleotide substitution rates. The effects of this bias can easily be misconstrued as the action of purifying selection at synonymous sites.

Chapter 3 is concerned with mutational variation in the murid genome. Nucleotide substitution rates in murid transposable elements were estimated. It was assumed that the majority of murid transposable elements were evolving neutrally and, therefore, that their molecular evolutionary rate was dictated by mutation alone. Under this assumption, variation in estimated element substitution rates reflects sampling and mutational variation only. The results indicate that greater mutational variation occurs along the length of a chromosome than between individual chromosomes, although the latter has been the primary focus in the literature. This result illustrates the importance of accounting for mutational variation in studies of selective constraint and sequence conservation.

In Chapter 4, the level of constraint in intergenic DNA adjacent to coding sequences and a moderate distance inside first introns was estimated in a

sample of 300 mouse-rat gene orthologues. The results suggested that whilst selective constraint in intergenic sequence adjacent to the start and stop codons is moderately high, this becomes statistically indistinguishable from zero within 4kb upstream/downstream of the first/last exon. Selective constraint in the 5′ end of the first intron was also found to be moderately high. Taking the contributions from noncoding sequence into account, it was estimated that the number of deleterious mutations occurring in murid noncoding DNA was approximately equal to that in protein-coding sequence.

Chapter 5 expands on the work done in Chapter 4. The assumption of neutral evolution in non-first introns was addressed by comparing their evolutionary rates with those in transposable elements. In addition the selective constraint in intergenic DNA immediately adjacent to genes with that found large distances from known genes was compared. The results showed that, when repetitive sequence is removed, the selective constraints in intergenic DNA are significantly different from zero. Furthermore, this constraint does not become indistinguishable from zero, even at large distances (50kb) from genic regions. The data also showed that a weak correlation between intron length and nucleotide substitution rate exists in murid non-first introns.

Contents

1. Introduction

1.1. Evolutionary Significance of Noncoding DNA

Until relatively recently, studies of molecular evolution in eukaryotic genomes have focused primarily on protein-coding genes. However, it has been apparent for some time that the fraction of the genome that is translated into protein comprises a relatively small proportion of the total genome length, in many eukaryotes. Few genomic features are more puzzling than the large amounts of noncoding sequence typical of many eukaryotic species because, whilst the function of a gene (i.e. to produce a protein) is relatively obvious, the purpose and significance of the majority of noncoding DNA is unclear.

The C-value Paradox and the History of Noncoding DNA

The discovery of noncoding DNA is closely linked to what became known as the "C-value paradox". The term "C-value", coined in the 1950s (Swift, 1950), originally referred to the apparent constancy in the amount of DNA per nucleus within the tissues of an organism. The observation that DNA content per cell was constant within, and therefore characteristic of, a species was first suggested in 1948 (Vendrely and Vendrely, 1948), and was seen as evidence that DNA, as opposed to protein, was the hereditary material. It was also quickly realised that substantial interspecific variation in C-value existed (Mirsky and Ris, 1951). The C-value paradox arose when it became evident that, although DNA was clearly the chemical basis of heredity, no obvious relationship existed between the amount of DNA per cell and organismal "complexity". Although based upon the somewhat nebulous concept of complexity, the C-value paradox generated a substantial degree of controversy. This was primarily due to the general assumption (e.g. Commoner, 1964) that C-value must be directly related to the species gene number which, in turn, must be correlated with complexity.

The resolution of the C-value paradox came with the accumulation of evidence that in many species genome size bears little relationship to the amount of protein-coding sequence. A number of factors were instrumental in this change

of perspective. Perhaps most influential was the discovery that, rather than a collection of distinct, unrelated genes, many eukaryotic genomes contained relatively enormous quantities of repeated sequences (Britten and Kohne, 1968). This became apparent from experiments that measured the time taken for separated strands of DNA from the same species to reassociate. It was quickly noticed that a certain proportion of strands taken from mammalian species reassociated very rapidly, even more swiftly than viral DNA, whilst another fraction reassociated far more slowly (Waring and Britten, 1966). This result indicated a high degree of self-similarity in the rapidly annealing fraction. In addition it became clear that, even within the nonrepetitive region of the genome, the proportion of sequence which actually codes for protein is relatively small (Lewin, 1975b). This conclusion was reached from observations that the precursor molecule of mRNA, dubbed heterogeneous nuclear RNA (hnRNA), tended to be larger, sometimes substantially so, than the mRNA molecule itself (Lewin, 1975a). The resolution of this conundrum came with the discovery that genes themselves were not continuous but split, and interspersed by noncoding regions, dubbed "introns" (Gilbert, 1978). Furthermore, introns appeared to be often much larger than the coding regions themselves. From the combined evidence for large amounts of repetitive, noncoding DNA, and genes split by further, apparently functionless, intragenic noncoding regions, it was clear that notion of a genome consisting solely of functional, protein-coding genes was incorrect. These early studies laid the foundations for the prevailing perception that, whilst some noncoding DNA in many species is likely to be functional, this portion is in the minority.

Transposable Elements

The discovery of large quantities of noncoding DNA, much of which appeared to lack any protein-coding function led to the hypothesis that eukaryotic genomes were primarily composed of essentially useless "junk". Although, in its original context, "junk" strictly referred to the nonfunctional pseudogenes that litter the genome (Ohno, 1972), this later became regularly used as a collective term for noncoding sequence. The notion of a genome that was composed of large regions which apparently had little effect on phenotype stimulated the development of theories of intragenomic parasitism and "selfish" DNA (Orgel and Crick, 1980; Doolittle and Sapienza, 1980). Specifically it was suggested that the origin of much noncoding sequence originated from the proliferation of "selfish"

replicators. The discovery of bacterial insertion sequences (Hirsch et al., 1972) and *Drosophila* P-elements (Engels and Preston, 1980) proved that both prokaryote and eukaryote genomes did indeed contain "selfish" sequences capable of autonomous replication and insertion into a "host" genome. Further weight was lent to this idea by the subsequent discovery of highly successful families of retrotransposing elements, such as Short and Long Interspersed Elements (SINEs and LINEs) within mammalian genomes (Singer, 1982).

Since their discovery, transposable elements have become the subject of much interest in molecular evolution. Whole genome sequencing has confirmed the success with which mammalian genomes have been invaded by self-replicating sequences, with transposable elements estimated to comprise at least 46%, 39.5%, 40%, and 31% of the human, mouse, rat and dog genomes, respectively (IHGSC, 2001; IMGSC, 2002; IRGSC, 2004; Kirkness et al., 2003). These figures are very likely to be underestimates, given the difficulty in identifying very old repetitive sequences which have experienced many nucleotide and insertion/deletion (indel) substitutions. It is clear that transposable elements have had an enormous influence upon the evolution of mammalian genome size. A more controversial debate is the impact of transposable element insertion upon organismal fitness and their relevance (or lack thereof) to evolution at the phenotypic level.

Although it was "repugnant" (Britten and Kohne, 1968) to early investigators that an abundance of DNA could exist without conferring some direct and immediate selective benefit, as yet, no conclusive evidence has been offered that transposable element insertion is beneficial in more than a minority of cases. There have been a number of reports of transposable elements that appear to have been co-opted for function post-insertion (Britten, 1997). However, the majority of these appear to have acquired function a substantial time after insertion and are unlikely to have been immediately functional. Comparative genomics studies have indicated that repetitive DNA is substantially underrepresented in datasets of conserved, putatively functional motifs (Margulies et al., 2003; Siepel et al., 2005) The recent discovery of an entire repeat family (MER121) which is highly conserved across the mammalian phylogeny (Kamal et al., 2006), appears to be the exception rather than the rule. Functional repeats, therefore, appear to represent a small fraction of the total number of transposable elements present in most mammalian genomes.

What is clear is that transposable element insertions, in common with most mutations, can have immediate deleterious effects on organismal fitness. These effects can generally occur in two ways. Firstly, elements may insert into functionally important DNA, often exonic or related to gene regulation, and disrupt normal gene function. This appears to be the case for a number of human genetic diseases, where insertion of an *Alu* element is likely to be the causative mutation (Deininger and Batzer, 1999). Secondly, ectopic recombination between homologous elements inserted into different genomic regions may cause deleterious chromosomal rearrangements. Again, examples of this are known from human genetic diseases (e.g. Lehrman et al., 1987; Vidal et al., 2002). From these examples, it would seem that transposable element insertion appears to impose a selective cost much more frequently than it confers a selective advantage upon the host organism.

Given apparent fitness detriment often imposed by transposable element activity, their sheer abundance in many eukaryotic genomes requires explanation. This is not as counterintuitive as it first seems. It has been realised for some time that it is quite possible for selfish replicators to spread to fixation in a sexual population, without necessarily conferring a selective advantage (Hickey, 1982). In contrast, asexual lineages are less likely to become colonised by self-replicators unless they have an opportunity to spread inter-genomically by a some mechanism, such as horizontal transfer. In fact, even if element insertions are phenotypically deleterious, their ability to replicate faster than the host genome means they can still rise to appreciable frequencies in sexual populations (Hickey, 1982). The strength of selection against transposable element insertion will depend on how often they insert into functional regions and the size of the population in which they arise. It is worth noting that abundance of transposable elements in mammals and vertebrates is not reflected in many invertebrate species. In *Drosophila* the repetitive portion of the euchromatic genome is estimated at ~5% (Quesneville et al., 2005), although the total genomic fraction, including heterochromatin, is likely to be higher. The genome sequences of *Caenorhabditis elegans* and *C. briggsae* also harbour a somewhat lower fraction of repetitive sequence (16.5% and 22.4%, respectively; Stein et al. 2003) than in mammals. It is tempting to conclude that these differences reflect differences is the efficacy of natural selection to remove parasitic sequences between invertebrates and vertebrates, resulting from a reduction of effective population sizes in vertebrates. However, the situation is evidently not this

straightforward, given that some likely counterexamples, such as the brown mountain grasshopper (Bensasson et al., 2001), exist.

Whether the actual insertion of a transposable element is deleterious is not, however, expected to impose constraints on the rate at which that element accumulates nucleotide substitutions. With little evidence to suggest that motifs encoded by transposable elements become functional at a significant rate, it seems likely that inserted elements that drift to fixation in a population generally accumulate point substitutions at the neutral mutation rate.

Introns

The discovery that gene protein-coding sequences were not continuous but interrupted by regions of silent, noncoding DNA was entirely unexpected. Molecular cloning of genes such as mouse immunoglobulin (Brack and Tonegawa, 1977) and in the 28s rRNA gene of *Drosophila* (Glover and Hogness, 1977), revealed that the structure of eukaryote genes was more complex than had previously been thought. Specifically, it appeared that coding sequences within genes were frequently split by long stretches of intragenic, noncoding DNA. Since then, much work has focused upon the origins, persistence and potential functions of introns. In particular, two questions have dominated many evolutionary studies of introns: whether introns arose early or late in evolutionary history and what features of introns are responsible for their persistence in many modern day taxa.

Introns can be classified into three groups. Group I and II introns are found in some bacteria and eukaryotic organelles and are self-splicing. In contrast, spliceosomal introns are removed from the pre-mRNA molecule by a complex of RNAs and proteins known as the spliceosome. The "introns early", "introns late" debate arose from observations of the phylogenetic distribution of spliceosomal introns. Notably, this class of introns are entirely absent from eubacteria, archaebacteria and many single-celled protists and yet ubiquitous in multicellular eukaryotes (Lynch and Richardson, 2002). Both hypotheses are essentially derived from alternative interpretations of this pattern, and relate to the importance of introns in early gene evolution.

The "introns early" hypothesis suggests that introns are very ancient and were

subsequently lost from prokaryotes and many single-celled eukaryotes, perhaps as a result of a streamlining of their genome for greater functional or transcriptional efficiency (Gilbert, 1978; Doolittle, 1978; Blake, 1978). This hypothesis suggests that introns originally existed to facilitate the modular assembly of genes from early exon fragments via "exon shuffling". This hypothesis predicts that introns arose interspersed between ancient mini-proteins to create "whole" genes. Some studies have suggested that there is little relationship between exons and protein domains, as would be expected if modern genes were initially assembled from small, exonic fragments (Stoltzfus et al., 1994; Stoltzfus, 2004). Others have, however, suggested exactly the opposite (Liu and Grigoriev, 2004). A further prediction of the "introns early" view is that, because ancient genes were assembled from distinct modules, intron-exon boundaries should correspond to codon boundaries more often than expected by chance. A statistically significant proportion of introns in ancient genes do indeed appear to be phase zero (i.e. the intron-exon boundary coincides with a codon-codon boundary) (De Souza et al., 1998). However, this pattern could also be produced by preferential insertion of new introns within specific dinucleotides which, in turn, are favoured at codon boundaries due to codon bias (Qiu et al., 2004; Belshaw and Bensasson, 2006; Ruvinsky et al., 2005).

The alternative view of intron origins, the "introns late" hypothesis, holds that spliceosomal introns arose only in eukaryotes. This theory proposes that gene segmentation in eukaryotic genes arose via random insertion of sequences, possibly ancient transposable elements, into early, continuous eukaryotic genes (Orgel and Crick, 1980). Furthermore, the machinery to excise introns from protein-coding genes is suggested to have been derived from the self-splicing mechanism of group II introns found in some prokaryotes (Cavalier-Smith, 1991; Logsdon and Palmer, 1994). The implication is that introns were never present in the extant taxa which lack them. Support for the shared ancestry of spliceosomal introns and prokaryotic introns comes from observed similarities in the self-splicing mechanism of group II introns and the spliceosome (Lynch and Richardson, 2002; Belshaw and Bensasson, 2006).

The more extreme opinions regarding the origins of spliceosomal introns have moderated since the initiation of the debate, over two decades ago. For example, with the discovery of group I and II introns in prokaryotes, it seems reasonable to conclude that that introns, in one form or another, were present very early on

in evolutionary history. Likewise, however, the balance of evidence is in favour of "late" (i.e. eukaryotic) origin of spliceosomal introns given that none of over 100 sequenced prokaryotic genomes has shown any evidence of spliceosomal introns (Lynch and Richardson, 2002). However, although it does appear that spliceosomal introns originated in eukaryotes, their origins appear to stretch back to the very deepest branches in the eukaryotic lineage (Nixon et al., 2002). By definition, the "introns early" and "introns late" hypotheses make entirely opposing predictions about the rates of intron loss and gain over evolutionary timescales and this remains a further unresolved point. The "introns early" prediction of a massive loss of spliceosomal introns across all lineages leading to extant prokaryotic species is, however, considerably less parsimonious than the alternative, "introns late" view. The controversy surrounding rates of intron loss and gain has yet to be resolved satisfactorily, however, with multiple analyses apparently demonstrating both large-scale intron losses (Roy et al., 2003) and gains (Wolf et al., 2001; Qiu et al., 2004) in a variety of taxa.

There have been a number of proposed selective advantages of introns which could have lead to their retention. The original version of the "introns early" hypothesis suggested that primordial introns existed in ancient taxa as a mechanism of creation and evolution of early proteins via exon-shuffling (Gilbert, 1978). This theory postulates that the exon-intron structure of eukaryotic genes arose as a mechanism whereby proteins with diverse functional roles could be created via the duplication and rearrangement of exons. A similar hypothesis suggests that introns play an important role in generating multiple protein isoforms from the same gene via alternative splicing. Selective advantage of introns due to a role in alternative splicing would provide support for the "introns late" hypothesis, however, given that alternative splicing is only known from eukaryotes.

Introns have also been proposed to function in mRNA error-correction via nonsense-mediated decay (NMD; Lynch and Kewalramani 2003). NMD is a surveillance mechanism that checks mRNA for splicing errors, such as premature stop codons and faulty open reading frames, which could lead to nonfunctional or harmful proteins (Hentze and Kulozik, 1999). When such aberrant mRNA molecules are identified they are degraded before they can be exported to the cytoplasm. Introns are thought to play a crucial role in NMD, by functioning as markers for the true stop codon. In eukaryotes, each exon-intron boundary

is labelled with a complex of proteins before translation. NMD functions by detecting stop codons that occur prior to the last of these labels, which is generally located before the true termination codon, and therefore likely to result from transcriptional error or a mutant allele. NMD is also known from yeast, although here the mechanism relies upon downstream sequence elements embedded within exonic, rather than intronic, DNA (Gonzalez et al., 2001).

Introns may also play important roles in other aspects of gene regulation. For example, work in *Xenopus* has produced some evidence that mRNA produced from intron-containing genes are more efficiently exported to the cytoplasm than identical mRNA transcribed from cDNAs (Luo and Reed, 1999).

A final proposed utility of introns is as recombination modifiers between exons. Insertion of introns between different parts of a gene necessarily increases the rate of recombination between these different gene regions. This has been suggested as potentially beneficial as it could reduce the effects of Hill-Robertson interference and thus increase the efficacy of natural selection upon weakly selected alleles (Comeron and Kreitman, 2002). This hypothesis arose from observations that longer introns tend to be located in regions of low recombination in *Drosophila* (Comeron and Kreitman, 2000). However, intron length appears also to be related to other variables, for example gene expression level (Castillo-Davis et al., 2002). In addition, correlations between intron length and recombination could just as easily arise from an inability to remove deleterious insertion mutations in low recombination rate regions (Carvalho and Clark, 1999). Finally, it also appears that there is a strong positive correlation between intron length and selective constraint in *Drosophila* (Haddrill et al., 2005; Halligan and Keightley, accepted), a pattern which is also evident in intergenic DNA (Urrutia and Hurst, 2003). This would suggest that longer introns tend to function in capacities more vital than recombination modification.

The ubiquity of spliceosomal introns in eukaryotes need not necessarily result from a definite selective benefit of their presence, however. An alternative view of intron evolution proposes that long-term changes in effective population size are the major determinant of the phylogenetic distribution of introns. Lynch (2002) suggests that, although intron creation may initially be deleterious, the selection coefficient against them may be small enough that introns can become established in taxa of sufficiently low effective population size. This theory predicts a higher number and larger sizes of introns in taxa with lower

effective population size, a prediction that is, to some extent, borne out in reality (Lynch and Conery, 2003). However, even if introns did initially become fixed by chance and not as a result of some definite selective benefit, there is evidence to suggest that at least some have secondarily acquired useful functions. This is supported by evidence that longer introns are more highly selectively constrained (Haddrill et al., 2005; Halligan and Keightley, accepted) in *Drosophila*. It is likely that at least some of these constrained regions function in gene regulation, and this is supported by the observation that more developmentally complex genes tend to harbour longer introns (Nelson et al., 2004). In addition, the bimodal distribution of intron lengths characteristic of many species suggests that some selective advantage is bestowed by introns (Yu et al., 2002).

Evolutionary Conservation and Selective Constraint in Noncoding DNA

Despite the development and popularity of "selfish" or "junk" DNA hypotheses to explain the existence and proliferation of noncoding DNA, it has been known for a relatively long time that at least some noncoding sequence must be functional. For example, it was clear from the late sixties onwards that the sites controlling transcription initiation, collectively known as the promoter, do not themselves code for a protein molecule in either prokaryotes or eukaryotes (e.g. Epstein and Beckwith, 1968). The discovery of the TATA box in eukaryotes, and its analogue the Pribnow box in prokaryotes, highlighted one clear function of noncoding DNA. With the development of DNA sequencing technology, it was soon possible to undertake quantitative, cross-species analyses of protein-coding genes and the noncoding DNA that surrounds them. Almost immediately it was recognised that this approach could prove invaluable in distinguishing functional regions from true junk, and the term "phylogenetic footprinting" was coined (Tagle et al., 1988). Under the null hypothesis, based upon Kimura's neutral theory, that nucleotide substitutions in nonfunctional DNA occur at the mutation rate (Kimura, 1983), by definition anything that deviates significantly from this null (in particular by evolving at a rate substantially lower than expected under neutral evolution) is a good candidate for functional DNA. Cross-species comparisons of orthologous genes, employed in early studies, provided the first clear evidence that DNA in introns (Hayashida and Miyata, 1983; Emorine et al., 1983) and untranslated regions (UTRs; Cowan et al. 1983; Yaffe et al. 1985) could contain functional, selectively constrained motifs.

One of the first studies to compare multiple gene orthologues across a wide variety of vertebrate groups revealed strong sequence conservation in both UTRs and the intergenic DNA that flanked them (Duret et al., 1993). This study was one of the first to demonstrate that evolutionarily conserved, putatively functional noncoding DNA was not isolated to a few specific cases, but instead appeared to be a general characteristic of vertebrate genome evolution. Within a relatively short space of time, comparison of the syntenic sequence surrounding orthologous immune system genes in mouse and human (Koop and Hood, 1994) provided among the first evidence that putatively functional noncoding DNA could also be found deep within intergenic regions. Work on mammalian globin genes (Gumucio et al., 1996), provided evidence that conserved regions of noncoding DNA could indeed function in gene regulation, an important experimental validation of the utility of comparative methods. This experimental result was further supported by an analysis of sequence surrounding the Bruton's tyrosine kinase (BTK) gene in human and mouse which indicated that conserved blocks of noncoding DNA adjacent to the first exon of the BTK gene were involved in the regulation of its expression (Oeltjen et al., 1997). The utility of very wide phylogenetic comparisons in identifying functional noncoding DNA became evident with the availability of sequence data from the Japanese pufferfish, *Fugu rubripes*, (e.g. Aparicio et al., 1995).

With the initiation of large scale sequencing projects, truly genome scale comparative analyses became possible in mammals. Early comparative analysis of the sequence derived from draft versions of the human and mouse genomes suggested that noncoding sequence consisted of mosaics of constrained and randomly drifting sequence in intronic (Jareborg et al., 1999) and intergenic (Shabalina et al., 2001) DNA. Constrained, putatively functional blocks in both intronic and intergenic DNA also emerged from comparisons of orthologous noncoding regions between *Drosophila melanogaster* and *D. virilis*(Bergman and Kreitman, 2001).

The first, true genome wide comparative analysis of mammalian (human-mouse) noncoding DNA came with the publication of the draft mouse genome sequence, and suggested that protein-coding sequences only account for approximately one-fifth of the total amount of each species' genome under purifying selection (IMGSC, 2002). The majority of comparative analyses on a genomic scale since have also uncovered significant amounts of noncoding

DNA which appear to be conserved across large phylogenetic distances (Dermitzakis et al., 2002, 2003; Thomas et al., 2003; Margulies et al., 2003; Bejerano et al., 2004; Siepel et al., 2005). A typical example of this type of genome wide survey is that of Dermitzakis et al. (2002). Their initial study compared 33.5 megabases (Mb) from the long arm of human chromosome 21 with syntenic sequence in mouse. Alignment of these regions revealed a high frequency of well-conserved, ungapped sequences located primarily in the Giemsa-dark, gene-poor region of chromosome 21. The majority (2,262) of these conserved non-genic (CNGs) sequences appeared not to match any known genes on chromosome 21. CNGs appeared to be, on average, both smaller (approximately 150bp in length compared with average chromosome 21 exon length of 270bp), and over twice as numerous as known exons. The authors provide some evidence, based primarily on sequence analysis of the CNGs, that they are unlike known protein-coding genes.

One criticism which can be directed at this study is that, because the identification of CNGs was based on extremely large numbers of comparisons of substantial amounts of sequence, whilst some CNGs may indeed be functional, a sizeable proportion could simply reflect the large sampling bias introduced by this procedure and not true evolutionary conservation. The authors addressed this issue to a certain extent by demonstrating that at least some proportion of the original set of CNGs identified between human and mouse were conserved enough to be amplified by polymerase chain reaction (PCR) in species as widely diverged as African elephant and platypus (Dermitzakis et al., 2003). In addition, this study demonstrated that those CNGs which could be located in various outgroup species were significantly more highly conserved than either coding sequence or noncoding RNA genes. The issue of sampling bias in CNGs was also addressed by Keightley et al. (2005a), who located the original human-mouse CNGs in two ingroup species (chimp and rat). One of the advantages of this approach is that any significant conservation of CNGs in either of these two lineages is nearly independent of the original sampling procedure. The results of this study indicated that selective constraint in CNGs, as calibrated by nearby, assumed neutrally evolving sequence, was substantial and the contribution of CNGs to the total number of selectively constrained bases in the genome was considerably larger than estimated in protein-coding genes. The extent of sequence conservation on human chromosome 21 was also independently verified by high-density oligonucleotide array analysis (Frazer et al., 2001). This study

used microarrays to compare 25-mers derived from human sequence data with orthologous dog and mouse DNA. The results suggested that ~40% of the human elements that appear to be evolutionarily conserved in mouse and dog were not similar to known exonic sequences.

Similar analyses of larger datasets, across multiple lineages have suggested that evolutionarily conserved blocks are a reasonably common feature of mammalian noncoding DNA. Analysis of sequence orthologous to a region of human chromosome 7 across 12 species by Thomas et al. (2003) identified 1194 elements, conserved across all 12 species, of which the majority (~70%) mapped to noncoding regions. This study is a significant methodological departure from that of Dermitzakis et al. (2002, 2003), in that the identification of conserved regions was confined to a well annotated region of known orthology in multiple species. A similar analysis of the same dataset using two alternate statistical frameworks by Margulies et al. (2003) confirmed the results of Thomas et al. (2003). This study provided some evidence that conserved noncoding regions could perhaps function as noncoding RNA genes or transcription factor binding sites, although these conclusions were based entirely upon predictive models rather than experimental verification. The study of Bejerano et al. (2004) revealed a small number of motifs that are 100% conserved between human, mouse and rat genomes. In addition, many of these regions also appear to be almost entirely unchanged in chicken. Of the 481 elements identified in this study, the authors estimate that just over half show little evidence of being transcribed in any of the species studied, based on comparison of their conserved regions with existing ESTs and mRNAs. The results of Siepel et al. (2005) described a similar pattern to these previous studies. However, this study investigated evolutionary conservation across a much wider diversity of species, from vertebrates, through a variety of *Drosophila* and *Caenorhabditis* species and yeast. Comparison of vertebrate genomes again revealed substantial conservation in noncoding regions, with a similar pattern evident in *Drosophila*. However, Siepel et al. (2005) note that the proportion of conserved regions located in noncoding DNA decreases with decreasing evolutionary "complexity" (from vertebrates, fruitflies, nematodes to yeast). This is not an altogether surprising result, given that the average quantity of noncoding DNA in each species group also decreases in the same direction. In addition, a major complication with drawing reasonable inference about the evolution of noncoding DNA from these results is that the evolutionary divergence between the members of each species

group varies dramatically from group to group. Thus, if noncoding DNA is generally less highly selectively constrained than coding sequence, as seems to be the case, then, by definition, proportionately fewer conserved noncoding regions will tend to be found with increasing evolutionary distance. Nonetheless, the results of this study are confirmation that sequence conservation within the noncoding regions of a large variety of species genomes is extensive.

Recent studies have also suggested that selective constraint (as distinct from sequence conservation) in *Drosophila* noncoding DNA is far more extensive than previously thought. Both Andolfatto (2005) and Halligan and Keightley (accepted) reported unexpectedly high selective constraints within *Drosophila* intergenic DNA indicating that sizeable amounts of noncoding DNA in fruitflies are functional. In particular the results of Halligan and Keightley (accepted) suggest that almost half the noncoding sites which are shared ancestrally between *Drosophila melanogaster* and *simulans* have some biological function. The scale of evolutionary constraint of noncoding DNA in fruitflies appears to exceed that observed in many vertebrate species to date.

One of the most notable features of many conserved noncoding regions is their genomic location. Dermitzakis et al. (2004) reported that the location of CNGs uncovered in a previous human-mouse comparison (Dermitzakis et al., 2002) appears to bear no direct relationship to the location of the nearest protein-coding sequence, and CNGs, instead, tend to be randomly distributed within intergenic regions. Although Dermitzakis et al. (2004) conclude that this is evidence that CNGs are unlikely to be *cis*-regulatory regions the spatial relationship between such elements and the genes they regulate is still unclear. The results of Dermitzakis et al. (2004) do not concur with those of Thomas et al. (2003) who estimate that their multi-species conserved sequences (MCSs) are preferentially located within 1kb upstream of the transcription start site. This may, however, be a feature of the ten genes in the region studied. Similarly the "ultraconserved" regions in humans uncovered by Bejerano et al. (2004) indicated that those elements which showed no evidence of protein-coding function appeared to cluster around transcription factors and developmental genes. Thus, the relationship of conserved noncoding sequences to nearby genes may depend on gene function.

For the moment at least, the nature and function of many conserved elements in

noncoding DNA remains unknown. This is mostly due to the large experimental effort required to ascertain the function of even a small number of candidate regions. However, a number of functional roles for noncoding DNA have been suggested. Perhaps the most well-characterised of these is the regulation of gene expression. There are many potential means by which gene expression could be controlled, and no doubt some that remain to be discovered. However, we can identify those elements whose functional roles are clear.

In eukaryotes transcription of DNA to form pre-mRNA is carried out by RNA polymerase II (Pol II). Regulating this process are a collection of transcription factors (e.g. TFIIA, TFIIB), proteins that bind to the DNA at specific sites. In order for transcription to be regulated correctly these transcription factors must be able to bind correctly to the DNA at a specific site. It is generally accepted that at least some conserved elements in noncoding DNA function as such binding sites. Among the more well understood of these sites are promoters. Promoters consist of a variety of motifs which enable the initiation of transcription of DNA by Pol II. Promoters consist of a "core" promoter, which often contains the TATA box, as well a variety of other proximal and distal elements. Promoters are critical to correct gene transcription and are, therefore, likely to be conserved over evolutionary time. In addition to promoters, a variety of other DNA-level elements, are recognised to be important in gene transcription control, including enhancers and silencers. Enhancer elements bind to trans-acting factors and elevate the transcription level of a gene whilst silencers perform the opposite function in a similar way by binding to repressors which decrease the rate of transcription of a gene. These DNA level regulatory elements can vary substantially in location, although promoters tend to be located upstream of a gene. Additional, less well-known elements, thought to limit the range of action of regulatory regions, known as insulators, are known to exist (Bell et al., 2001) and are also likely to be important, functional noncoding regions.

Some work has already attempted to establish whether conserved regions in noncoding DNA could include promoters, enhancers and silencers. An early example of this would include the experimental validation of two enhancers of the Hoxb-1 gene in mice which are also conserved in chicken and *Fugu* (Marshall et al., 1994). However, this study focused upon experimental validation of putative enhancers in transgenic mice *prior* to confirming their

evolutionary conservation in two other species. In contrast, the majority of more recent studies have adopted the reverse approach and used computational prediction to select candidate regions for further experimental work. The study of Loots et al. (2000) is a good example of the utility this method. Here the authors compared approximately 1Mb of orthologous, noncoding sequence between human and mouse and identified 90 conserved noncoding regions. When tested in transgenic mice and zebrafish, the largest of these blocks appeared to be involved in the regulation of three interleukin genes. Ghanem et al. (2003) followed a similar approach but searched for regions that were conserved in five vertebrate species. All regions conserved across all five species showed evidence of enhancer activity. A larger scale scan of human-mouse noncoding DNA surrounding the DACH gene by Nobrega et al. (2003) also produced evidence that a small proportion of the 1098 conserved noncoding regions they uncovered acted as long-range enhancers of the DACH gene. Although the labour intensive nature of much experimental work precludes the analyses of genome wide datasets, the scale of experimental validation is growing. More recently Woolfe et al. (2005) used a human-*Fugu* comparison to uncover 23 conserved noncoding sequences which show significant enhancer activity in four developmental genes. However, despite the apparent abundance of experimental confirmation of enhancer activity in conserved noncoding elements, it should be noted that there is significant ascertainment bias in the experimental work described. The studies of Marshall et al. (1994); Ghanem et al. (2003); Nobrega et al. (2003); Woolfe et al. (2005) specifically tested for enhancer activity of conserved noncoding regions, but did not test any alternative functional hypotheses. Thus, although the proportion of conserved noncoding regions which function in DNA based gene regulation is clearly not zero, the true proportion remains for the moment, almost completely unknown.

Another potential function of conserved noncoding DNA is as matrix or scaffold attachment regions (M/SARs). These sites anchor DNA sequence to the fibres of the chromosomal scaffold. They are thought to harbour AT rich motifs, replication origins and may also include transcription factor binding sites (Boulikas, 1993). Given that some 100000 M/SARs are thought to exist in the mammalian genome (Singh et al., 1997) these regions could potentially account for a sizeable proportion of conserved regions in mammalian noncoding DNA. As yet, however, this possibility has been addressed by only a single study. Glazko et al. (2003) estimated the fraction of noncoding DNA conserved between

human and mouse that was part of a predicted M/SAR at 11%. Although these results have yet to be replicated and were based upon predicted, rather than experimentally confirmed M/SARs, such regions could potentially explain at least some evolutionary conservation in noncoding sequence.

Perhaps the most interesting, and mysterious, functional category which could account for large proportions of conservation in noncoding DNA are RNA genes or ncRNAs. ncRNAs are RNAs which function without being translated into protein. Well-known examples of RNA genes would include transfer and ribosomal RNA (tRNA and rRNA). Because of their fundamentally important role in protein translation, both tRNA and rRNA have been known for a relatively long period of time. However, more recently other ncRNAs have been shown to perform an extremely diverse range of functions (see Storz, 2002). In addition, microarray studies of gene expression on human chromosomes 21 and 22 have revealed substantial transcription of genomic regions outside known and predicted protein-coding genes (Kapranov et al., 2002). Further transcriptomic analysis of these human chromosomes has suggested that, although large numbers of transcription factor binding sites are evident, roughly equal numbers of these sites appear to be related to protein-coding genes and ncRNAs (Cawley et al., 2004). This apparent excess of transcriptional activity is not confined to humans but also appears to be the case in *Drosophila* (Stolc et al., 2004) and yeast (Havilio et al., 2005). Some of this excess is likely to be due to transcription of ncRNAs. However it is also not unrealistic that much of what is transcribed in these genomes is functionless (Brosius, 2005) as many different cellular mechanisms, such as transcription of retrotransposed transposable elements, create noncoding RNAs without indicating any real function (Eddy, 2002). Despite the attempts of Dermitzakis et al. (2003) to eliminate ncRNAs as a potential explanation for human-mouse CNGs, it is unclear from their results how much conservation in noncoding DNA can be really explained by RNA genes. One of the reasons for this is that ncRNAs are inherently difficult to predict by computational methods as they lack the consistent statistical anchors (such as an open reading frame) that are used to predict protein-coding genes. With these experimental and computational difficulties hampering the discovery of new RNA genes, it is unclear at the present time what proportion of conservation in noncoding DNA could be explained by ncRNAs.

Some of the above studies have relied on computational comparative methods

to identify sequences conserved between species. Close identity between two sequences by itself does not, however, confirm the orthology of conserved noncoding sequences between groups. Although the probability that two nonorthologous sequences of a reasonable length being highly similar is small, this can become an issue when large numbers of comparisons are made to identify such sequences, such as the large-scale BLAST comparisons used by Dermitzakis et al. (2002, 2003). This is not a problem for studies which have located conserved blocks within whole-genome alignments of multiple species (e.g. Siepel et al., 2005). Additionally, whilst sequence conservation is often a signature of selective forces, accepting conservation *ipso facto* as proof of function is unrealistic. One of the most difficult objectives in any attempt to locate functional blocks via comparative analysis is the definition of a robust null hypothesis, although more rigorous statistical models are being developed (Siepel et al., 2005). A fuller understanding of the pattern of mutational variability across the genome is fundamental to the success of comparative genomics. Our knowledge of such variability, however, is still incomplete even though this could explain at least some observed conservation (Clark, 2001). Finally, although the studies described above have investigated the level of sequence conservation and selective constraint in noncoding DNA in a variety of species, it is important to note that whilst the two are related they are not the same.

Selective Constraint and Deleterious Mutation

The definition of selective constraint is rooted in the neutral theory of molecular evolution which will be discussed briefly. The original neutral theory described evolution at the molecular level in terms of two classes of mutation : those that are neutral, and those that are deleterious (Kimura, 1968). Kimura's original definition of a neutral mutation was one for which:

$$|s| \leq 1/(2N_e)$$

where s is the selection coefficient of the mutant heterozygote and N_e is the effective population size. Kimura's assertion was that the majority of new mutations in a population fall into the neutral category. This suggestion was provoked by two related experimental observations. Firstly, data from primate haemoglobins and a number of other mammalian genes showed that the rate of

amino acid substitution was of the order of one per diploid genome every two years. Following calculations by Haldane (Haldane, 1957), Kimura suggested that this rate of substitution was much too high for each substitution to be driven to fixation by natural selection. The essence of the Kimura's argument was that, in order to sustain this rate of fixation by selection alone, the substitutional load or "cost of natural selection" was enormous. In effect this meant that for a population to remain a constant size and experience fixation of one selectively driven allele every two years required a fertility excess of very fit individuals (those that have the selectively advantageous allele at many loci) that was impossibly large ($\sim 10^{78}$), incompatible with known mammalian family sizes (Kimura and Ohta, 1971). Kimura therefore suggested that the majority of substitutions consisted of mutants which were effectively invisible to natural selection and, therefore, fixed by the stochastic process of random genetic drift. Secondly, early molecular studies had also revealed substantial protein polymorphism in natural populations (Lewontin and Hubby, 1966). Kimura also proposed that this observation was more parsimoniously explained by substantial numbers of segregating neutral alleles, rather than some form of balancing selection, such as overdominance.

Since its proposal, the neutral theory has provoked much debate, and come under criticism for a number of reasons. Among the most damning of these are that the load calculations which provide one of the cornerstones of evidence for the neutral theory rely upon unreasonable assumptions such as multiplicative fitnesses across loci, which are likely to be biologically unrealistic. In addition, the suggestion of family sizes of the order of 10^{78} relates to extreme individuals possessing the selectively favoured allele at all loci. However, as has been pointed out (Ewens, 2004), it is extremely unlikely that this individual will ever exist. A further problem was Kimura's assumption that nearly all nucleotides in the genome are functional (Kimura, 1968), which is evidently not the case. Correcting Kimura's original estimate of the genomic rate of amino acid substitution with a more realistic estimate of the number of human genes (for example) unsurprisingly leads to a considerably lower rate (1 amino acid substitution every $\sim$ 600 years; Nei 2005). Despite these criticisms the neutral theory provides a valuable null hypothesis for the detection of natural selection in molecular data and it is this framework upon which the definition of selective constraint is based.

Selective constraint is defined as the fraction of mutations at a locus which are deleterious enough to be removed by purifying selection (Kimura, 1983; Ohta, 1992). Under the assumptions of the neutral theory that adaptive mutations arise with negligible frequency, selective constraint can be simply defined as $1 - f_0$, where f_0 is the fraction of neutral mutations. However, the criteria for what actually constitutes a deleterious mutation, according to the neutral theory, are dependent upon the population in which it arises. More specifically, mutations with a certain (negative) selection coefficient may arise in populations of very large effective size which are swiftly removed by selection. Mutations with the same selection coefficient arising in very small populations may rise to appreciable frequency before becoming lost or fixed, due to the reduced efficacy of natural selection in such populations to remove them. In terms of selective constraint, those new mutations which are classified as deleterious are those for which the selection coefficient is negative and the product $|2N_e s|$, is greater than one. Thus, even if it were the case that the selection coefficients of new mutations in all taxa were equal, selective constraints upon new mutations would vary from species to species, according to differences in effective population size.

1.2. Estimating Nucleotide Substitution Rates and Selective Constraint

The work in this thesis depends on a number assumptions regarding the estimation of the rate of nucleotide substitution. The most important is that, in two sister species, sites can be identified which are descended from the same single base in the last common ancestor. This criterion of orthology is crucial to any comparative study. In this respect at least, the mice and rats are suited to the study of molecular evolution. Because the evolutionary divergence between the two species is comparatively low, identification and subsequent alignment of putatively orthologous sequences is reasonably reliable.

In addition to this, accurate estimation of nucleotide substitution rates relies on the ability to account for the superimposition of substitutions which have occurred at the same site, known as "multiple hits". Various models, of varying degrees of complexity, have been suggested to correct for this problem (Jukes and Cantor, 1969; Kimura, 1980; Gojobori et al., 1982; Tajima and Nei, 1984; Hasegawa et al., 1985; Tamura and Nei, 1993; Yang, 1994). Throughout

this thesis, unless explicitly stated, the Tamura and Nei (1993) method has been used to correct for multiple hits. In general these models assume a Markov process of nucleotide substitution in a DNA sequence and require a number of simplifying assumptions regarding the nature of molecular evolution. The first, and perhaps most important, assumption is that of stationarity. In a molecular evolutionary context, assuming stationarity is equivalent to the assumption that nucleotide frequencies are at equilibrium and unchanging. This assumption is clearly violated in many cases during mammalian evolution. For example, multiple studies have shown that the base composition of the mammalian genome is changing (Duret et al., 2002; Belle et al., 2004). The rate of genomic compositional change appears to be quite slow in recent mammalian lineages, but swifter earlier on in the mammalian phylogeny (Belle et al., 2004). It therefore seems unlikely that substantial compositional shift has occurred since the split between mice and rats and that nonstationary nucleotide frequencies have unduly affected the results in this thesis. No attempt has been made to correct for this aspect of molecular evolution throughout the course of this thesis. A further important assumption of the common models used to correct for multiple hits, including those used in this thesis, is independence of nucleotide sites along a DNA sequence. Again, this assumption is clearly violated in vertebrate genomes, most notably by context-dependent hypermutation at the methylated CpG dinucleotide. This issue has been addressed by attempting to exclude such substitutions, as much as possible, from estimates of nucleotide substitution rates (*see* Chapter 2).

Throughout this thesis various adaptations of the method of Kondrashov and Crow (1993) have been used to estimate selective constraint in murid noncoding DNA. This method effectively relies on the comparison of rates of molecular evolution in a sequence that is presumed to be evolving neutrally with that suspected to be under some level of selective constraint. This method is based upon two assumptions. Firstly, it is assumed that all mutations that occur can be divided into two classes: neutral and deleterious. It can be seen that this assumption has its roots in the neutral theory of molecular evolution. Given the uncertainty regarding the distribution of mutational effects it is, however, still unclear whether this assumption is valid, although it is generally accepted that most new mutations in functional regions of the genome will be deleterious. Adaptive substitutions will inflate the observed evolutionary rate and thus lead to an underestimate of selective constraint. The second assumption

of the method of Kondrashov and Crow (1993) is that deleterious mutations will not contribute to evolution. As discussed above, the evolutionary prospects for a new deleterious mutation depend upon the population in which that mutation has arisen. However, we can clarify this by stating that this method will give an estimate of the frequency of occurrence of new, deleterious mutations with selection coefficients greater than the reciprocal of twice the effective population size (Kondrashov and Crow, 1993). Given that these assumptions hold, selective constraint (C) at a locus can be estimated as follows:

$$C = \frac{\mu_N - \mu_O}{\mu_N}$$

where μ_N is the neutral mutation rate and μ_O is the observed evolutionary rate at that locus. Thus, if a locus is evolving entirely neutrally, μ_N will equal μ_O and constraint is zero. Constraint is therefore estimated as the fraction of "missing" substitutions which are assumed to have been deleterious, and removed by natural selection.

1.3. Aims of this study

In this study, an investigation of the magnitude of, and spatial variation in, selective constraint in murid noncoding DNA was carried using a comparative genomics approach. In addition, work to test some of the underlying assumptions of the method of estimating constraint was attempted. By definition, selective constraint can provide information about the rate of deleterious mutation, an important parameter in evolutionary theory, providing we have a reliable estimate of the base mutation rate. Selective constraint can also help determine the evolutionary significance of at least some noncoding DNA.

In Chapter 2 some problems associated with CpG hypermutation in the estimation of nucleotide substitution rates are addressed. CpG sites are the most mutable sites in the mammalian genome, due to the high rate of deamination of methylated cytosine to thyamine within a CpG dinucleotide. For this reason, studies of molecular evolution in mammals have often divided estimated nucleotide substitution rates into those inferred to have occurred within and outside sites that were CpG in the ancestral sequence. However, a common method of assigning ancestral CpG state has the potential to introduce serious bias into substitution rate estimation, particularly when rates are compared across sites with different CpG frequencies, such as fourfold degenerate and

noncoding sites. This is of interest because comparisons of substitution rates at these sites have recently been used to infer purifying selection at mammalian synonymous sites (Hellmann et al., 2003; ICGSC, 2005). More generally, this is relevant because of the potential for the introduction of serious bias into the estimation of nucleotide substitution rates. We address the issue of misassignment using simulations of a simple two-branch phylogeny evolved under a CpG hypermutation model.

Chapter 3 concerns variation in mutation rate within the murid genome. The mutation rate is thought to vary across mammalian genomes (Wolfe et al., 1989; Ellegren et al., 2003). This has important consequences for "phylogenetic footprinting" as the expectation of sequence conservation across evolutionary time by chance will vary with genomic region. If mutation rates vary substantially across relatively short scales, this could explain some of the block-like patterns of sequence conservation observed in many comparative studies of mammalian genomes (e.g. Dermitzakis et al., 2002). In addition, the scale of mutational variation is an important factor in determining the adequacy of putatively neutrally evolving sequence for the calibration of selective constraint. Finally, although there have been many estimates of between chromosome mutational variation in mammals (Lercher et al., 2001; Taylor et al., 2006), there has been little work to estimate the level of mutational variation within chromosomes. This is potentially important as it can provide information about the relative importance of the some of the processes that drive mutation. A dataset of transposable elements that were inserted into the last common ancestor of mice and rats was collected and their nucleotide substitution rates estimated. Under the assumption that the majority of the ancestral repeats are evolving neutrally, any variation in nucleotide substitution rate was assumed to result from sampling or mutational variation. A simple linear model was used to partition mutational variation into inter- and intra-chromosomal components.

Chapters 4 and 5 deal with selective constraint in murid noncoding DNA. Although, protein-coding sequence is known to be highly selectively constrained (Eyre-Walker and Keightley, 1999) in mammals as well as *Drosophila* (Keightley and Eyre-Walker, 2000), little is known about selective constraint in mammalian noncoding DNA. Selective constraint of noncoding DNA is important for a variety of reasons. Firstly, variation in the level of selective constraint can indicate where functionally important regions of DNA tend to

be located. Secondly, selective constraint can be used to measure the rate of deleterious mutation. The genomic deleterious mutation rate U is a crucial parameter in population genetics. The value of U has a particular bearing upon one theory for the evolution of sexual reproduction. The "Mutational Deterministic" (MD) hypothesis suggests that sexually reproducing species can overcome the twofold disadvantage of sex compared to asexually reproducing species if U is greater than one (Kondrashov, 1988). The additional requirement of the MD hypothesis is synergistic epistasis between deleterious mutations, such that additional deleterious mutations have proportionally greater effect than expected under simple additivity. Providing these two criteria are met, then purging the genome of deleterious mutation could provide a long-term advantage for sexual reproduction. Previous studies for a variety of taxa have relied on estimates of U within coding sequence alone (Keightley and Eyre-Walker, 2000). However, it is likely that at least some noncoding DNA is functional and under purifying selection. Thus, deleterious mutations occurring in noncoding DNA could contribute substantially to the total genomic deleterious mutation rate. The aim of both these chapters was to quantify the level of selective constraint in murid introns and intergenic regions in order to resolve the evolutionary significance of noncoding DNA.

2. Ancestral CpG assignment at fourfold synonymous and noncoding sites

2.1. Introduction

It is widely accepted that the methylated form of cytosine (5-methylcytosine or 5mC, Figure 2.1) is hypermutable (Bird, 1980). 5mC is formed by DNA methyltransferase operating on a cytosine occurring immediately 5′ of a guanine. Such sites are typically referred to as "CpG" dinucleotides, the "p" referring to the phosphate group between the cytosine and guanine. One effect of methylation is to increase the rate of spontaneous deamination of 5mC to form thymine. It has been estimated that transitions in the methylated CpG dinucleotide occur 8 (Arndt et al., 2003b), 10 (Siepel and Haussler, 2004) or even 16 times faster (Lunter and Hein, 2004) than non-CpG transitions. A smaller elevation of the rate of transversion mutation at the CpG dinucleotide has also been observed (Hess et al., 1994; Blake et al., 1992; Siepel and Haussler, 2004). Elevation of CpG mutation rate may also vary throughout the genome (Arndt et al., 2005). A recent analysis of substitution rates in transposable elements has suggested that mutation rates in methylated CpG dinucleotides in mammals underwent an abrupt elevation at sometime around the mammalian radiation ($\sim$ 90Myr; Arndt et al. 2003b). One possible explanation of this phenomenon is an increase in the level of germline methylation in mammals, possibly in response to invasion by highly replicating transposable elements (Yoder et al., 1997).

As a result of the elevation of mutation rates at methylated CpG dinucleotides, studies of molecular evolution have frequently attempted to separate the estimation of CpG and non CpG substitution rates. In alignments of two sister species, this has typically been attempted by separating observed substitutions into those that are inferred to have occurred within and outside a CpG dinucleotide. In pairwise alignments a site which was part of a CpG in the ancestral sequence is defined as any nucleotide which occurs either within or opposite a CpG in either sequence. Likewise, a site which was non CpG ancestrally is defined as any nucleotide which does not occur within or opposite

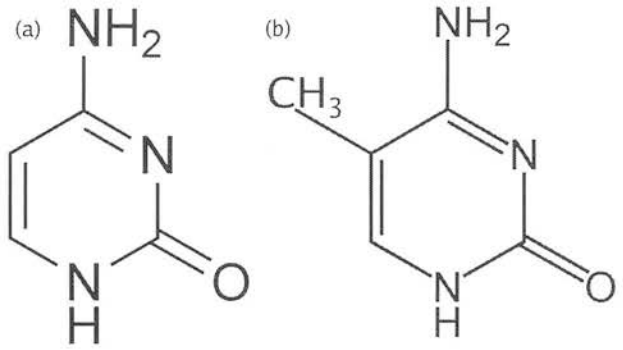

Figure 2.1.: The chemical structure of cytosine (a) and 5-methylcytosine (b).

a CpG dinucleotide in either species. This is subsequently referred to as "CpG assignment" or "non CpG assignment", respectively. This method has previously been employed in comparisons of the substitution rate at synonymous sites with that in noncoding DNA, such as introns (Ebersberger et al., 2002; Hellmann et al., 2003; Rosenberg et al., 2003; Subramanian and Kumar, 2003; ICGSC, 2005). Some studies have suggested that, whilst the overall rate of substitution is higher at synonymous sites, when rates are separated into CpG and non-CpG substitutions using CpG or non CpG assignment, both rates are lower, sometimes substantially, than those estimated in noncoding DNA (Hellmann et al., 2003; ICGSC, 2005). This observation has been inferred to result from the action of negative or purifying selection at synonymous sites.

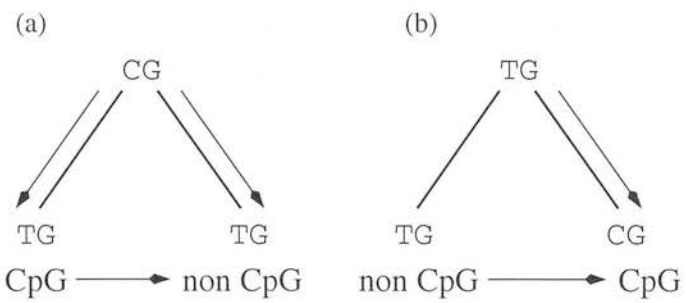

Figure 2.2.: Examples of CpG and non CpG misassignment. Arrows denote a single nucleotide substitution.

However, a problem with CpG/non CpG assignment arises from the questionable assumption that ancestral CpG dinucleotides can be reliably identified by their presence or absence in two extant, derived sequences. Clearly, when using CpG/non CpG assignment, the potential exists for misclassification of true CpG changes as non CpG and *vice versa*. This would not present a

serious issue if the chance of misclassification of CpG changes was identical or very similar to the chance of misclassification of non CpG changes. However, it can be seen that, in order for an ancestral CpG to be misclassified as non CpG in two derived sequences, at least two nucleotide substitutions must occur, one in each copy of the CpG in each lineage (Figure 2.2 a). This is hereafter referred to to as "CpG misassignment". In contrast, for at least some ancestral non CpG dinucleotides to be classified as CpG in two derived sequences, only a single nucleotide substitution (to "C" or "G", depending on context) is required (Figure 2.2 b). This is hereafter referred to as "non CpG misassignment". Given the difference in the minimal number of substitutions required to produce CpG versus non CpG misassignment, it is likely that these biases will have differential impacts upon the estimation of nucleotide substitution rate, depending on a variety of parameters including the level of CpG hypermutability, the evolutionary distance between the two species and the sequence base composition. This is of interest because, as mentioned, CpG and non CpG assignment has previously been used to estimate CpG and non CpG nucleotide substitution rates. In particular, this method has been used to compare evolutionary rates at synonymous and noncoding sites, between which there is known, and substantial, compositional variation in at least some mammalian species (Figure 2.3). At fourfold synonymous sites compositional differences are exaggerated because of the structure of the genetic code (the probability that a randomly chosen fourfold synonymous site is preceded by a "C" is 0.5). In addition, the sites flanking fourfold synonymous sites are typically under strong purifying selection.

The purpose of this chapter is, therefore, to investigate the magnitude of bias introduced by CpG/non CpG assignment into the estimation of nucleotide substitution rates and how this bias varies with base composition and level of CpG hypermutability over different evolutionary timescales. In particular it was investigated whether biases introduced by CpG/non CpG assignment can explain differences in the estimated evolutionary rates at synonymous and noncoding sites. It is also shown that removing "CpG-prone" sites (sites preceded by "C" or followed by "G") is a reliable way of removing the effects of CpG mutation. Throughout this chapter the analysis of coding sequence is restricted to fourfold degenerate sites. Of the 576 (64×9) possible codon mutations, 129 are synonymous. Of these, the majority (96) can only occur at a fourfold synonymous site. In addition many recent analyses of substitution

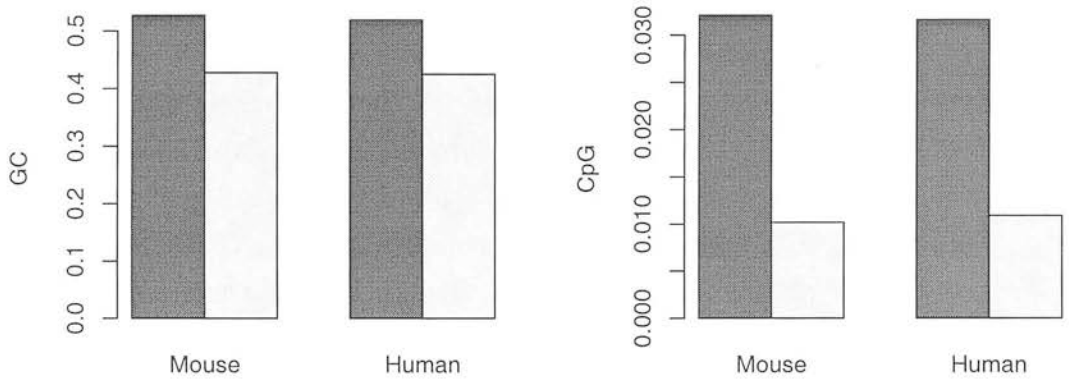

Figure 2.3.: The proportion of a sample of real mouse and human coding (dark bars) and noncoding (light bars) sequences made up of by "G" and "C" bases (a) and CpG dinucleotides (b). Dark bars represent coding sequence, light bars represent intronic sequence. Mouse sequence data were the same as analysed in Chapter 5. Human sequence data were provided by Dr. Peter Keightley.

rates at synonymous sites have been restricted to fourfold sites (e.g. ICGSC, 2005). Thus, whilst these results do not extend to all synonymous sites, they are relevant for the vast majority of synonymous substitutions that can occur.

In this study a simulation approach was adopted to model sequence evolution with CpG hypermutability. Although dinucleotide mutation models (Arndt et al., 2003a; Lunter and Hein, 2004; Hwang and Green, 2004) have previously been used to model the evolution of dinucleotide frequencies (Sved and Bird, 1990), this approach is inadequate as it fails to take account of "edge effects" which arise because dinucleotides overlap.

2.2. Materials & Methods

The analysis was restricted to simple, two-branch phylogenies. Two sequences were copied from a single, ancestor sequence and evolved. A simple mutation model was implemented where transitions occurred twice as frequently as transversions, and CpG mutations occurred with higher probability than non CpG mutations. Higher or lower levels of transition/transversion bias had minimal effects upon the results (results not shown).

Ancestral sequences in were created in two ways. The first approach involved simulating random sequences with a variety of GC contents, ranging from 0.0-1.0. These are hereafter referred to as "artificial ancestral" sequences.

The base composition of real mammalian sequence data is complex, however, and difficult to simulate without losing important features. For example the frequency of the CpG dinucleotide in mammalian DNA is much lower than would be expected from the product of its constituent base frequencies. Features such as this deficiency have a dramatic impact on the rate of molecular evolution and were, therefore, important to simulate. To achieve this, real coding and noncoding data were used as ancestral sequences. The real sequence dataset consisted of all mouse coding sequence collected in Chapter 5 and intronic sequences from the same dataset of approximately the same length ($\sim$ 8Mb). These are hereafter referred to as "real ancestral" sequences. In simulations where "real ancestral" sequences were used, all sequence data was concatenated into a single sequence and evolved.

The simulation of coding sequences was as follows. In the "artificial ancestral" simulations, random sequences of a given GC content at fourfold degenerate sites were simulated, whilst in the "real ancestral" simulations, real, unaugmented mouse coding sequence was used as the ancestral sequence. In both cases, nonsynonymous substitutions were rejected with a certain probability, which was defined as selective constraint of the sequence. A true fourfold nucleotide substitution was defined as any substitution which occurred at a fourfold degenerate site. To obtain the observed number of fourfold nucleotide substitutions, only those fourfold codons which coded for the same amino acid in both derived sequences and which had experienced a single synonymous change were classified as ancestrally fourfold. In all comparisons between the estimated and true rate of nucleotide substitution in phylogenies derived from "real ancestral" sequences, a constraint of 1 was simulated. All simulations were written in C. The computer code of the simulation program is presented in Appendix A.

2.3. Results

Dynamics and approximate equilibrium frequency of the CpG dinucleotide at fourfold and noncoding sites

It was first addressed whether there are substantial differences in the evolutionary dynamics of the CpG dinucleotide frequency at fourfold synonymous and noncoding sites. The expected equilibrium frequency of CpG in a sequence has been studied previously using a dinucleotide mutation model (Sved and Bird, 1990). However, as previously mentioned, this approach is flawed in that it assumes that dinucleotides evolve independently of one another, when in fact dinucleotides in a sequence form an unbroken series of overlapping pairs. Here this question was investigated using simulations. For simplicity, sequences were generated so that the starting CpG frequency at fourfold sites was equivalent to that in noncoding sequence (0.0625, the expected CpG frequency in a random sequence of GC content 0.5). Throughout, CpG frequency is defined as the number of observed CpG dinucleotides in a sequence divided by the total number of possible dinucleotides in a sequence ($n - 1$, where n is the number of nucleotides in the sequence). At fourfold sites, CpG frequency was defined as the total number of CpG dinucleotides in which one or other of the constituent bases was fourfold degenerate, divided by the total number of dinucleotides possible at fourfold sites ($2n$, where n is the number of fourfold degenerate sites).

The results show that, as expected, the frequency of the CpG dinucleotide decays with evolutionary time, measured in mean number of substitutions per site (Figure 2.4). In addition it is clear that, starting at the same initial composition, even after a comparatively small period of evolutionary time, the frequency of the CpG dinucleotide is higher at fourfold sites than in unconstrained noncoding DNA. Furthermore, the rate of decay of the CpG dinucleotide at fourfold sites depends on the level of selective constraint at the flanking sites, with a more gradual rate of decay in more highly constrained coding sequence.

The approximate equilibrium frequency of the CpG dinucleotide in coding and noncoding sequence was also investigated (Figure 2.5). At equilibrium, the number of CpG dinucleotides "created" by non CpG mutation over any period of time equals the number of CpG dinucleotides "destroyed" by CpG mutation in the same time period. Approximate equilibrium frequencies were estimated after evolving a sequence such that each site had, on average, experienced two

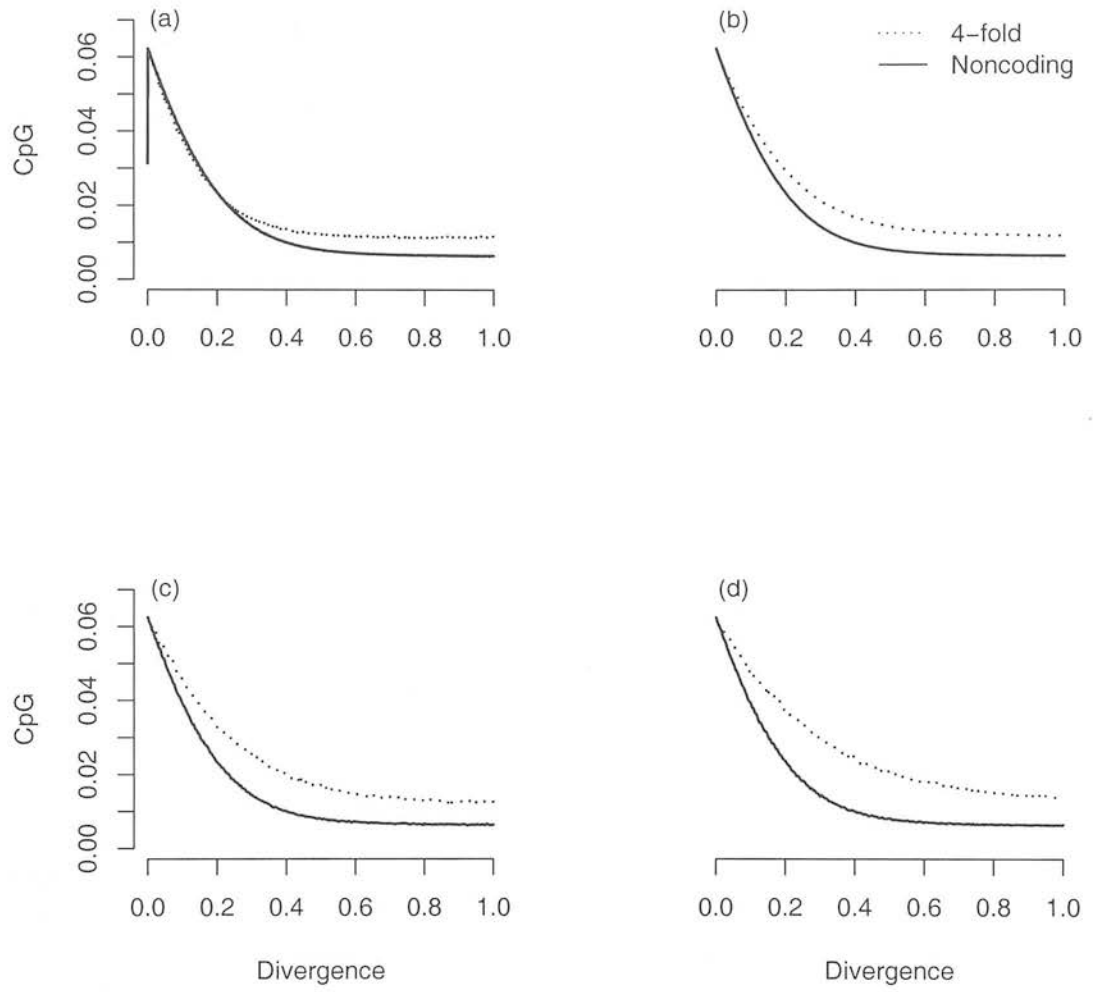

Figure 2.4.: Change in frequency of the CpG dinucleotide with sequence divergence (measured in numbers of nucleotide substitutions per site) for coding (dotted lines) and noncoding (solid lines) derived from "artificial ancestral" sequence. Plots are shown for coding sequence selective constraints of 0.0 (a), 0.5 (b) 0.75 (c) and 1 (d). Each line represents 250 data points, each of which was estimated from a single, 3Mb sequence, evolved under 10-fold CpG hypermutability.

nucleotide substitutions. The plots of CpG decay in Figure 2.4 suggest that after this divergence, the frequency of CpG in both coding and noncoding sequence was near to equilibrium. The results show that, under the same mutational regime, fourfold sites always tend towards a higher CpG frequency at equilibrium than noncoding DNA. It is clear that the major determinant of the equilibrium frequency of CpG is the level of hypermutability. The elevated equilibrium frequency at fourfold sites in an unconstrained coding sequence

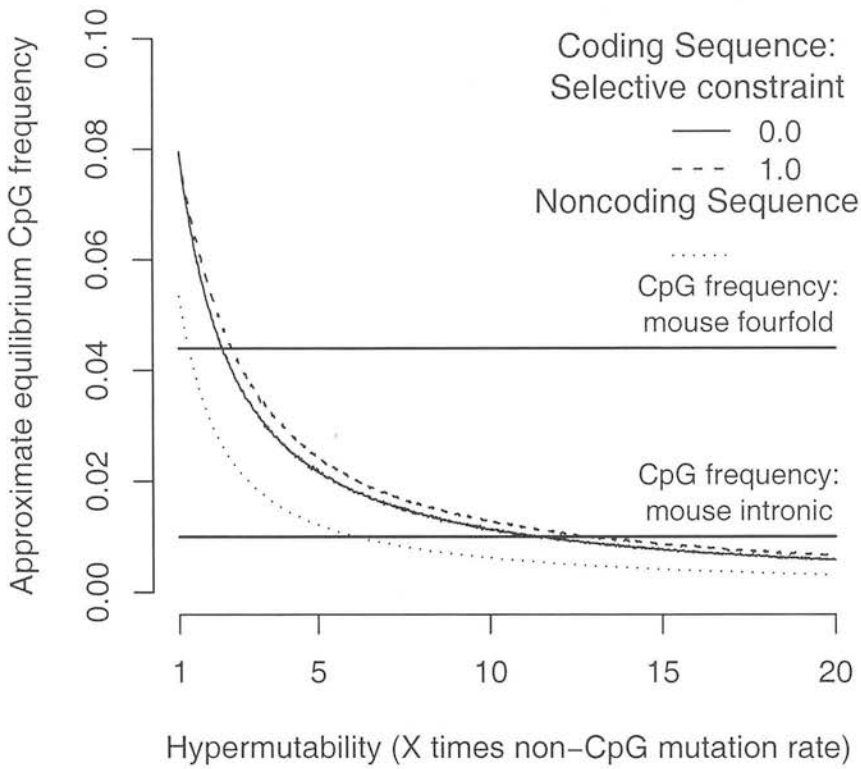

Figure 2.5.: Approximate equilibrium frequencies of the CpG dinucleotide in coding and noncoding sequence across different levels of hypermutability. Approximate equilibria are given for a variety of coding sequence selective constraints and for neutrally evolved noncoding sequence. Horizontal, solid lines show the observed CpG frequency at mouse fourfold (upper line) and intronic (lower line) sites. Each line represents 100 data points, each of which was estimated from a single, 3Mb sequence, evolved until each site had experienced two nucleotide substitutions, on average. Although, after this divergence on average $\exp^{-2} \approx 13\%$ of sites will not have experienced a substitution, this will not be the case at the constituent nucleotides of a CpG dinucleotide which will generally reach equilibrium much more swiftly than non CpG sites. The criterion of two nucleotide substitutions per sites was chosen by visual inspection of the plots in Figure 2.4.

demonstrates the bias introduced at fourfold sites by the structure of the genetic code. With non zero selective constraint the CpG frequency is marginally elevated, perhaps due to a selective preservation of nearest-neighbour sites at fourfold sites. In real sequence data, nonrandom usage of fourfold codons will inevitably further influence fourfold CpG frequency. At 10-fold hypermutability, the approximate equilibrium CpG frequency ranges from 0.011 to 0.013 at

fourfold sites (corresponding to selective constraint ranging from 0 to 1). In contrast, the approximate equilibrium frequency at noncoding sites evolved under the same mutational model is 0.006, almost half that at fourfold sites in an unconstrained coding sequence. These frequencies are substantially lower than those found at real murid fourfold (0.044) and intronic (0.010) sites. This disparity could result from a number of factors. It may be that CpG frequency is still equilibrating following the putative elevation of CpG hypermutability (Arndt et al., 2003b) after the mammalian radiation. Further complexities in the real murid mutational spectrum which were not simulated may exist which also impact upon CpG dinucleotide frequencies. Additionally, nonrandom amino acid usage will also influence the frequency of CpG at fourfold sites. Selection (perhaps for increased GC content) could also play a role in maintaining a higher CpG frequency than expected from mutational biases alone (Kondrashov et al., submitted). In addition to these factors, CpG frequency in real sequences will be influenced by varying mutation rates and levels of methylation. Biased gene conversion is also thought to influence variation in base composition across the human genome (Meunier and Duret, 2004). Finally, CpG frequencies in real sequence data could also be influenced by other context dependent effects, such as the UA avoidance which appears to be a feature of some vertebrate genomes (Ohno, 1988).

These results suggest that if initial coding sequence GC content is greater than or equal to the noncoding GC content (which is the case in the mouse and human genome), after a small period of evolutionary time under elevated CpG mutability, the frequency of the CpG dinucleotide at fourfold sites is always higher than noncoding sequence under the same mutational model. Furthermore, even if coding sequence GC content is initially lower than that of noncoding sequence, the results suggest that the equilibrium CpG frequency of fourfold sites is higher than that in noncoding sequence.

Bias in ancestral CpG and non CpG assignment - "artificial ancestral"

Having established that fourfold degenerate sites are inherently more likely to have a higher CpG frequency than noncoding DNA, it was necessary to investigate the impact of compositional variation upon ancestral CpG/non CpG assignment. Two branch phylogenies with a variety of initial GC contents were simulated and evolved. CpG and non CpG substitution rates were estimated

using CpG/non CpG assignment of the ancestral state. The results demonstrate two, reciprocal problems arising from this method of assignment of the ancestral CpG state.

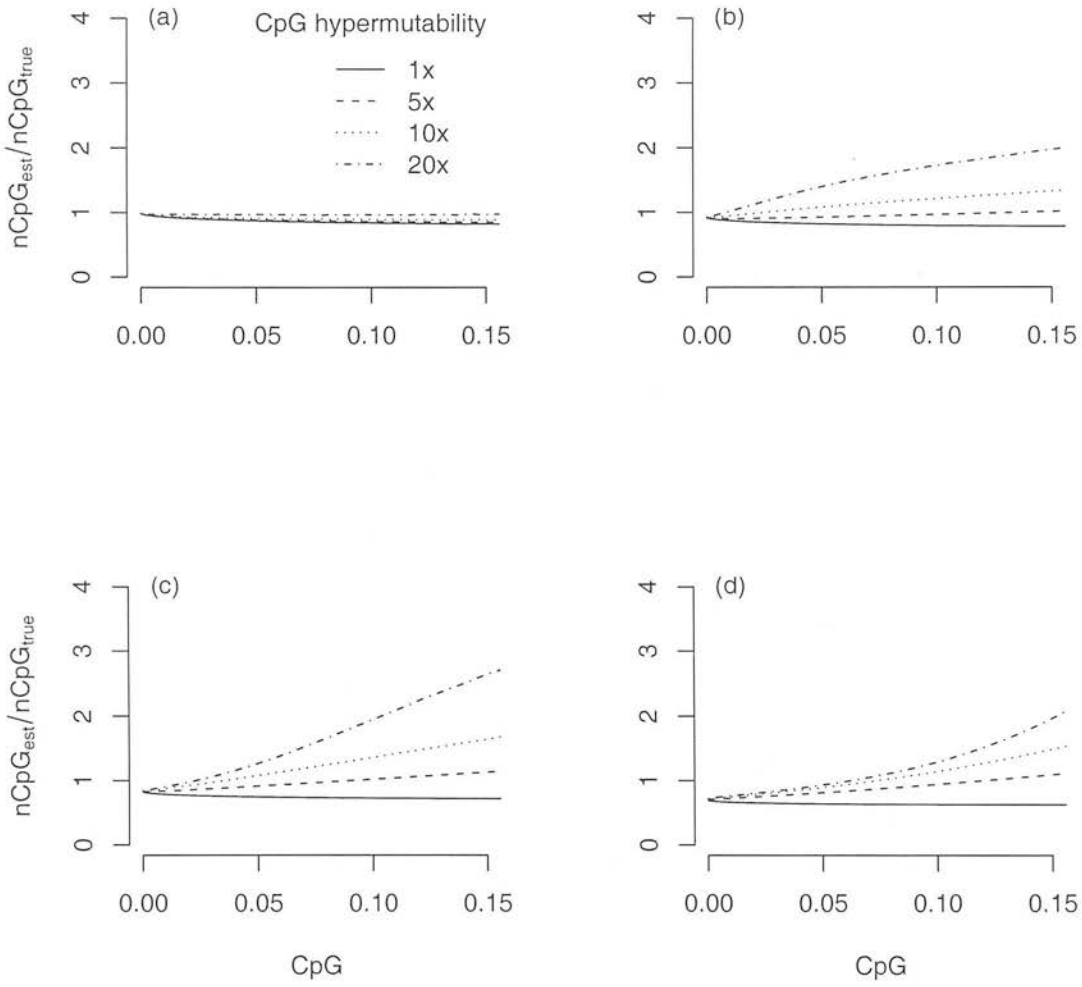

Figure 2.6.: Ratio of estimated ($nCpG_{est}$) to true ($nCpG_{true}$) non CpG differences across a variety of ancestral CpG frequencies for phylogenies derived from "artificial ancestral" sequences. Results are from trees with a total length of 0.01, 0.05, 0.1 and 0.5 (a,b,c and d respectively). Separate lines show results for different levels of hypermutability (no hypermutability, 5-fold, 10-fold and 20-fold hypermutability). Each line represents 101 data points, each of which represents 10 simulated replicates of a 1Mb sequence.

This first problem with CpG/non CpG assignment arises when the number of true CpG changes, misassigned as non CpG (*e.g* Figure 2.2 a), exceeds

the number of true non CpG changes, misassigned as CpG (*e.g* Figure 2.2 a). In this case the number of non CpG changes is overestimated. The results of simulations using "artificial ancestral" sequences demonstrate that the magnitude of this overestimation increases with increasing evolutionary distance and CpG hypermutability (Figure 2.6). This is because increasing values of both parameters increases the probability of the two or more CpG substitutions required to misinfer a CpG change as a non CpG change. These results also show that overestimation of the number of non CpG changes is problematic at relatively high ancestral CpG frequencies. For typical mammalian DNA the CpG frequency is quite low (e.g. mouse coding sequence: CpG frequency $\sim$ 0.032). In addition, these results demonstrate that overestimation of the number of non CpG changes is minimal across small evolutionary distances (1%). Thus, it is likely that an appreciable overestimation of the number of non CpG changes does not occur in comparisons between taxa similar to murids or hominids.

The second problem with CpG/non CpG assignment is the opposite of the first, namely when the number of true non CpG changes misassigned as CpG is greater than the number of true CpG changes misassigned as non CpG. In this scenario the number of true CpG changes is overestimated. As expected, the results of simulations demonstrate that the level of overestimation is inversely related to CpG hypermutability and evolutionary distance (Figure 2.7). In fact, the level of overestimation is most substantial when there is no hypermutability because, in this situation the highest of proportion of non CpG substitutions occur. This in turn, increases the number of non CpG changes which are misassigned as CpG. In addition, simulations also show that the overestimation is at its most extreme at low CpG frequencies. Thus, it appears that overestimation of the true number of CpG changes will be problematic for species in which the frequency of the CpG dinucleotide is low, and between minimally diverged sister species.

The relationship between the two misassignment problems described here and the frequency of CpG in the ancestral sequence is interpreted as follows. It has already been shown that the level of overestimation of the number of non CpG changes increases with increasing ancestral CpG frequency. In this case, as the number of non CpG sites decreases, the proportional impact of an excess of misassigned CpG changes becomes greater. Likewise, the results show that the level of overestimation of the number of CpG changes increases with decreasing ancestral CpG frequency. Here, as the number of CpG sites becomes smaller, the

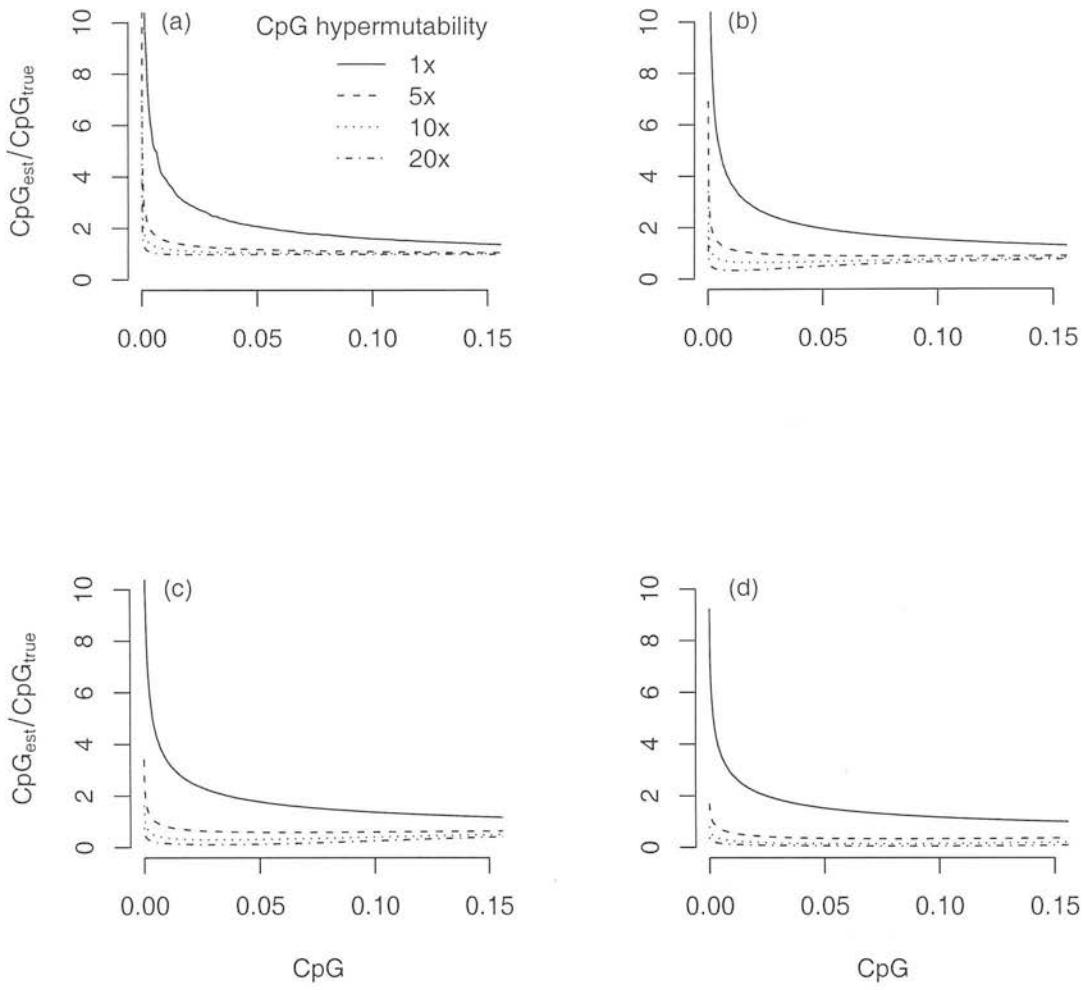

Figure 2.7.: Ratio of estimated (CpG_{est}) to true (CpG_{true}) CpG differences across a variety of ancestral CpG frequencies, for phylogenies derived from "artificial ancestral" sequences. Results are from trees of total length 0.01, 0.05, 0.1 and 0.5 (a,b,c and d respectively). Separate lines show results for different levels of hypermutability (no hypermutability, 5-fold, 10-fold and 20-fold hypermutability). Each line represents 101 data points, each of which represents 10 simulated replicates of a 1Mb sequence.

proportional impact of an excess of misassigned non CpG changes becomes larger.

These results suggest that the appropriate conditions exist in at least some mammalian species for the overestimation of the CpG substitution rate using CpG/ non CpG assignment to be problematic. This bias therefore was focused upon for further analysis. Naturally, as increasing numbers of true non CpG changes are misassigned as CpG, the estimated number of non CpG changes

also becomes biased downwards. A rescaling of the results in Figure 2.6 (a) is presented in Figure 2.8 and demonstrates this effect. With increasing evolutionary divergence, the number of any type of substitutional change is underestimated due to multiple hits, which have not been corrected for here. However, Figure 2.8 shows results for two minimally diverged sequences (1%), and so the number of multiple hits will be negligible. Thus, underestimation of the number of non CpG changes observed in Figure 2.8 is due almost entirely to the misassignment of true non CpG changes as CpG.

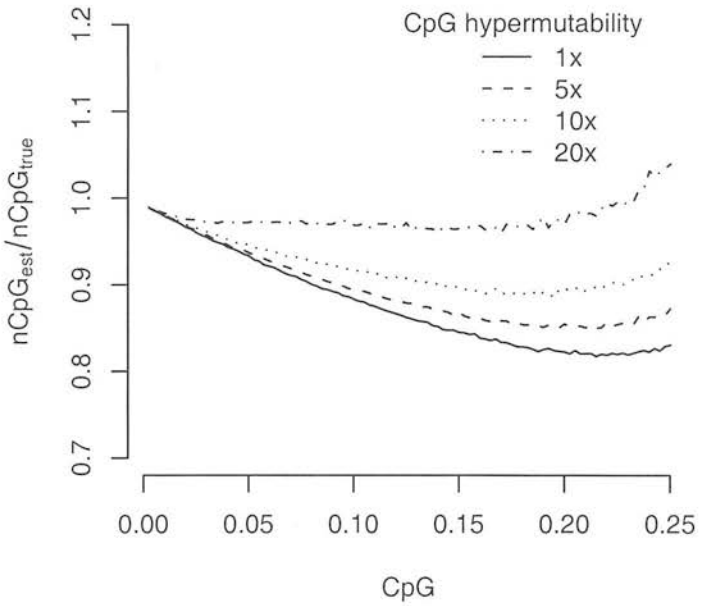

Figure 2.8.: Ratio of estimated ($nCpG_{est}$) to true ($nCpG_{true}$) non CpG differences across a variety of ancestral CpG frequencies for phylogenies derived from "artificial ancestral" sequences. Results are for a tree of total length 0.01. Separate lines show results for different levels of hypermutability (no hypermutability, 5-fold, 10-fold and 20-fold hypermutability). Each line represents 101 data points, each of which represents 10 simulated replicates of a 1Mb sequence.

Bias in ancestral CpG/non CpG assignment - "real ancestral"

Although the simulations using "artificial ancestral" sequences revealed the general relationship between base composition and the impact of CpG/non CpG misassignment, they are not entirely applicable to real sequence data. The reason

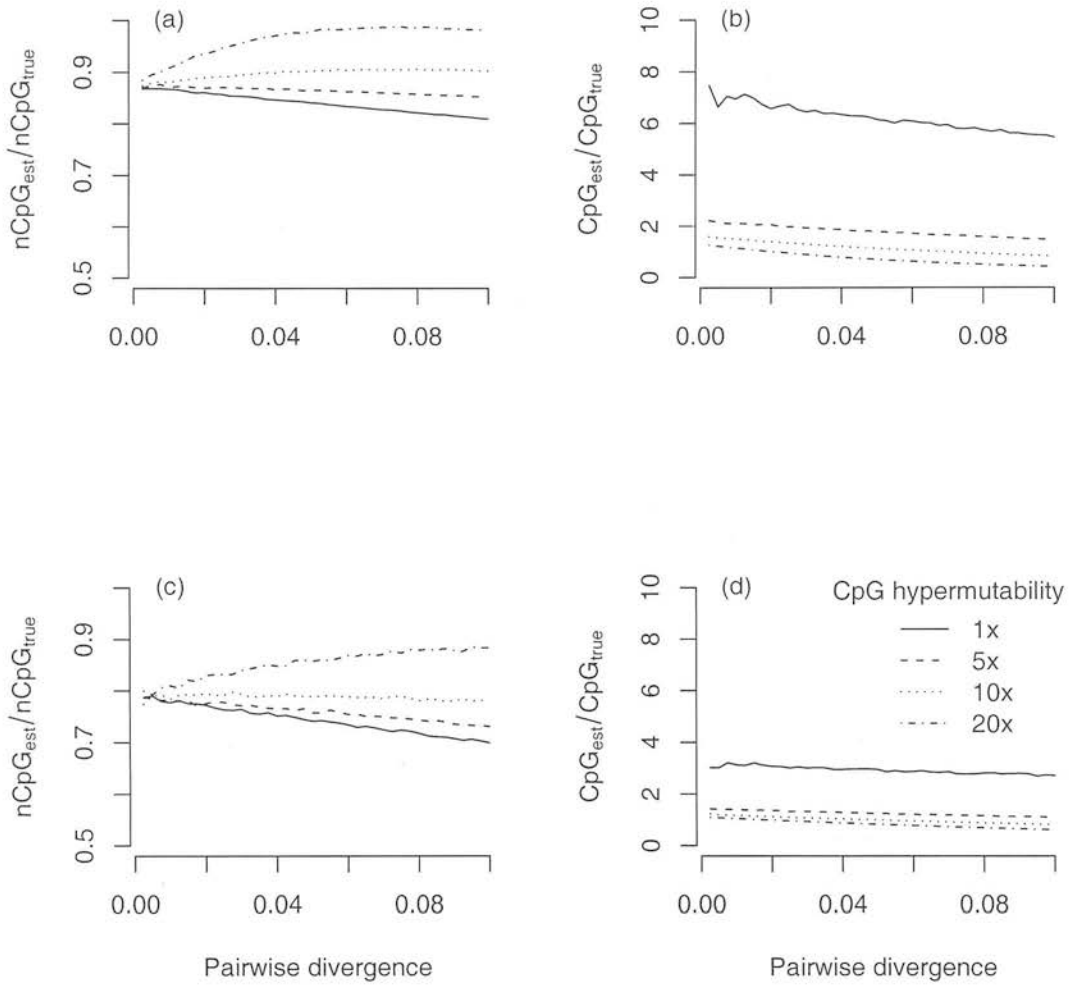

Figure 2.9.: Ratio of estimated (CpG_{est}) to true (CpG_{true}) CpG differences and estimated ($nCpG_{est}$) to true ($nCpG_{true}$) non CpG differences with increasing pairwise divergence, for phylogenies derived from "real ancestral" noncoding (a,b) and coding (c,d) ancestral sequences. Separate lines show results for different levels of hypermutability (no hypermutability, 5-fold, 10-fold and 20-fold hypermutability). Each line represents 40 data points, each of which was estimated from two simulated replicates of an $\sim$ 8Mb "real ancestral" coding or noncoding sequence.

for this is that, although the dinucleotide frequencies in a randomly generated sequence are determined by the product of the frequencies of their constituent nucleotides, as mentioned previously, this is rarely the case in real data. In mammalian sequences in particular the frequency of CpG is much lower than expected randomly, as a result of its hypermutability as well as other potential factors. Thus, many mammalian genomes are characterised by a moderate

GC content, but very low CpG frequency. Because of the evident importance of sequence composition, the impact of CpG/ non CpG misassignment was determined in sequence data that were compositionally more realistic. Given the significant compositional differences between coding and noncoding sequence, the effects of CpG/ non CpG misassignment upon coding and noncoding sequence were examined separately.

Non CpG and CpG substitution rates were estimated using CpG/non CpG assignment in phylogenies derived from "real ancestral" murid coding and noncoding sequence data. The results of this analysis are presented in Figure 2.9. It is apparent that the magnitude of assignment biases clearly differs between coding and noncoding DNA. As in the "artificial ancestral" data, overestimation of the number of CpG substitutions decreases with increasing hypermutability and divergence. Interestingly, however, the magnitude of overestimation is larger in noncoding DNA. For example, at 10-fold hypermutability and a pairwise divergence of 1%, the number of CpG substitutions in noncoding DNA is overestimated by $\sim$ 50%, whilst the equivalent figure is $\sim$ 17% at fourfold sites. Furthermore, whilst the number of non CpG substitutions is underestimated in both coding and noncoding DNA, the magnitude of this underestimation is greater at fourfold sites. For the same parameter combination as described above, the number of non CpG substitutions is underestimated by $\sim$ 12% in noncoding DNA, but by $\sim$ 20% at fourfold sites. The reason for these differences is simply that the CpG frequency differs between fourfold and noncoding sites, and that the numbers of misassigned sites make up different proportions of the total lost or gained from the CpG and non CpG classes. Again, it is clear that the underestimation of the number of non CpG substitutions is far greater than could be expected as a result of multiple hits alone. As an example, at the maximum divergence shown in Figure 2.9 (10%), the underestimation of the number of true non CpG changes expected due to multiple hits is $\sim$1%. This is smaller than the underestimation observed at 10% evolutionary divergence, which ranges from 19% to 2% in noncoding sequence, and from 30% to 12% at fourfold degenerate sites (ranges correspond to 0-fold to 20-fold hypermutability, respectively).

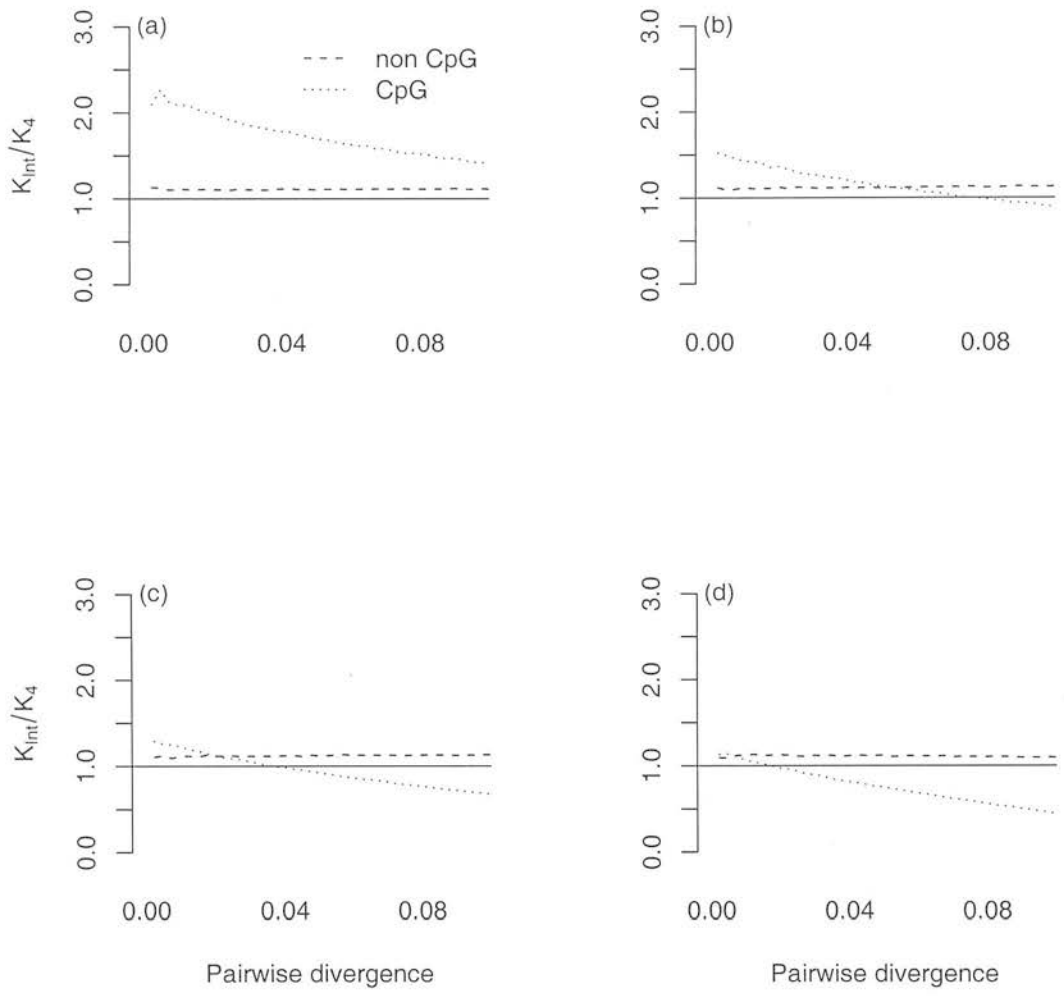

Figure 2.10.: Ratio of estimated substitution rate at fourfold degenerate and noncoding sites for phylogenies derived from "real ancestral" sequences. Rates are estimated at CpG and non CpG sites separately. Results are shown for four levels of CpG hypermutability : 1,5,10 and 20-fold (a,b,c and d respectively). Selective constraint of 1 was simulated at nonsynonymous coding sites. Each line represents 40 data points, each of which was estimated from two simulated replicates of an $\sim$ 8Mb "real ancestral" coding or noncoding sequence.

Artefactual elevation of CpG and non CpG substitution rate at noncoding sites

The results of the simulations described above have shown that (i) fourfold degenerate sites are predisposed towards a higher CpG frequency (Figures 2.4, 2.5) and (ii) that the ancestral frequency of CpG is a major determinant of the level CpG and non CpG misassignment (Figures 2.6, 2.7, 2.8). It therefore

seems that the level of CpG and non CpG misassignment will also differ between fourfold degenerate and noncoding sites. The effects of misassignment biases upon the estimation of the nucleotide substitution rate at these sites whilst implementing a standard correction for multiple hits (Tamura and Nei 1993) were therefore investigated.

Figure 2.10 shows the ratio of estimated substitution rates at noncoding (K_{Int}) and fourfold degenerate (K_4) sites, in data simulated using the "real ancestral" sequences. For small to moderate evolutionary distances, the results show that the ratio of K_{Int}/K_4 at non CpG sites is slightly above one across all divergences simulated. This appears to be relatively invariant to the level of CpG hypermutability. In contrast, the K_4/K_{Int} at CpG sites ratio varies widely with pairwise divergence and hypermutability level. These results reflect changes in the probability of a multiple hit at non CpG and CpG sites. Whilst the probability of a multiple hit at non CpG sites is unrelated to the level of hypermutability this will, by definition, have a large influence on the number of multiple hits at CpG sites.

The results of these simulations suggest that, at low divergences ($\leq$1%) and in molecular data that are compositionally similar to murid coding and noncoding sequence, the estimated CpG and non CpG substitution rates at noncoding sites will always exceed those estimated at fourfold sites, when CpG/ non CpG assignment is used. Thus, although both fourfold and noncoding sites were simulated to evolve free of any selective constraints, the substitution rates estimated at both site types could erroneously imply purifying selection at synonymous sites. The probability that estimated fourfold CpG and non CpG substitution rates are both less than the equivalent noncoding substitution rates decreases with increasing CpG hypermutability and evolutionary distance. Even with 20-fold hypermutability, however, the estimated CpG substitution rate is ~6% higher at noncoding sites than at fourfold sites for a pairwise divergence of 1%. Likewise, with 20-fold hypermutability, the estimated non CpG substitution rate is ~12% higher at noncoding sites, across a pairwise divergence of 1%.

Non-CpG-prone assignment removes CpG effects

These results demonstrate that even apparently small differences in CpG frequency have a large impact on the accuracy of estimation of non CpG substitution rates. It was therefore tested whether excluding CpG-prone sites

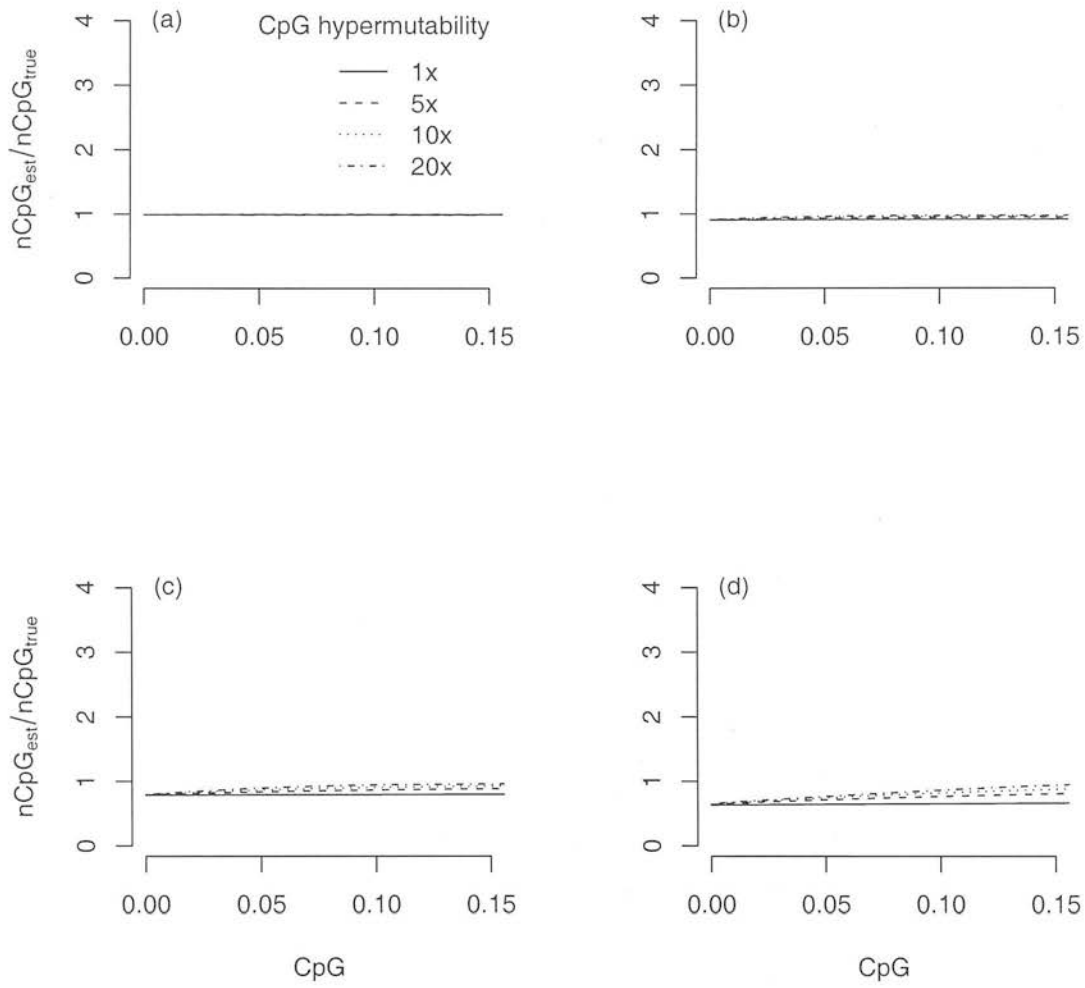

Figure 2.11.: Ratio of estimated ($nCpGp_{est}$) to true ($nCpGp_{true}$) non CpG-prone differences across a variety of ancestral CpG frequencies, in phylogenies derived from "artificial ancestral" sequences. Results are from trees of total length 0.01, 0.05, 0.1 and 0.5 (a,b,c and d respectively). Separate lines show results for different levels of hypermutability (no hypermutability, 5-fold, 10-fold and 20-fold hypermutability). Each line represents 101 data points, each of which represents 10 simulated replicates of a 1Mb sequence.

(those sites that are preceded by "C" and/or followed by "G") was an effective method of removing the influence of CpG mutation. The ratio of true to estimated non CpG-prone changes for a variety of simulated datasets is shown in Figure 2.11. The results show that the ratio of estimated to true numbers of non CpG-prone substitutions is close to unity even at extreme base compositions and large divergences. This compares favourably with the bias in the estimation

of the non CpG substitution rate in Figure 2.6.

2.4. Discussion

The results presented here have shown that a commonly used method of assigning nucleotide sites into CpG and non CpG categories systematically biases the estimation of nucleotide substitution rates at these sites over a wide range of base compositions, tree lengths and levels of hypermutability. Specifically, these results show that, across small evolutionary distances, CpG/non CpG assignment seriously upwardly biases the estimate of the number of CpG changes and downwardly biases the estimate of the number of non CpG changes. This occurs because of a simple artefact which means that, across small evolutionary distances, many more non CpG changes are misassigned as CpG changes than *vice versa*. As a result of this artefact, the differential base compositions of typical mammalian noncoding and fourfold degenerate sites are predisposed to misleading inferences about their relative rates of substitution. The net effect is that synonymous sites appear to be evolving more slowly than noncoding sites when rates are divided into those that have apparently occurred within and outside a CpG dinucleotide. It is important to note that the effect described here is not simply a problem of nonstationary CpG frequencies, as misassignment biases are still substantial when the number of CpG dinucleotides in both ancestral and derived sequences is unchanged.

Preliminary simulations showed that, for realistic parameter combinations, the equilibrium CpG frequency at fourfold sites is typically higher than unconstrained noncoding sites under the same mutational model (Figure 2.4). This is determined by two factors. Most importantly, elevated CpG frequency at fourfold degenerate sites simply results from the fact that randomly chosen fourfold sites are preceded by a "C" 50% of the time. Additionally, one of the nucleotides involved in the CpG dinucleotide at a fourfold site will generally be under purifying selection. Thus, fourfold CpG frequency is determined both by the mutation rate and the strength of selection. These results have some straightforward implications. Because of their elevated CpG frequency, fourfold sites will have a higher base mutation rate than noncoding sites. An explicit model of dinucleotide evolution or, failing this, efficient removal of CpG mutation, is therefore a prerequisite of any meaningful comparison of synonymous and

noncoding substitution rates. These compositional differences coupled with the problems in ancestral CpG assignment mean that, between closely related species (~1% divergence), it is almost inevitable that the estimated CpG and non CpG substitution rates at fourfold sites will be somewhat lower than those in noncoding sequence.

It is obvious from Figure 2.10 that CpG/non CpG assignment can seriously bias CpG substitution rate estimates at fourfold and noncoding sites. Despite the fact that both coding and noncoding sequences were evolved under identical mutational models, the estimated substitution rates only equal one another for very specific combinations of parameters. For the majority of parameter combinations examined it is unlikely that estimated fourfold and noncoding CpG substitution rates will be even approximately equal (whether greater or less than one another). Although the difference in estimated non CpG substitution rates between fourfold and noncoding sequence is slight, it is remarkably consistent across the pairwise divergences simulated.

The results presented here have implications for previous comparisons of the substitution rate at fourfold and noncoding sites in mammalian genomes. Misinference of purifying selection at fourfold sites due to the problems described here is apparently only possible across comparatively small evolutionary distances for a realistic level of hypermutability and the discussion is therefore restricted to situations where this is the case. Hellmann et al. (2003) present a comparison of substitution rates from fourfold and noncoding sites in human-chimp gene orthologues. Using CpG/non CpG assignment, they find that CpG and non CpG substitution rates at fourfold sites are ~40% and ~30% lower than in intronic DNA, respectively. The ICGSC (2005) also present a comparison of fourfold and intronic substitution rates and find that intronic CpG and non CpG substitution rates exceed those estimated at fourfold sites (by ~50% and ~30%, respectively). The results presented here suggest that at least some of this apparent reduction in fourfold substitution rates is artefactual. The extent of this effect depends on the level of hypermutability. Simulating 10-fold hypermutability in sequences derived from "real ancestral" coding and intronic DNA it was estimated that, across an evolutionary divergence of 1%, fourfold substitution CpG rates are ~18% lower than intronic CpG substitution rates as a result of misassignment alone. Likewise, estimated fourfold non CpG substitution rates are 10% lower than their intronic equivalents. At 5-fold hypermutability these percentages

increase to 30% and 11% respectively. Clearly, whilst there may be some real effect of selection on fourfold substitution rates it is probable that, in the analyses mentioned, it has been overestimated.

The prediction of higher rates at fourfold CpG and non CpG sites, compared with the equivalent intronic sites is not observed by Subramanian and Kumar (2003) however. These authors analyse a dataset of 83 human-chimp genes and observe no significant differences between the rate of non CpG substitution at fourfold sites and a variety of noncoding sequences. This is surprising because the simulations presented in this chapter suggest that, for a pairwise divergence and sequence base composition similar to that observed by Subramanian and Kumar (2003), both the non CpG and CpG nucleotide substitution rates will be artefactually deflated at fourfold sites, relative to noncoding sites, as a result of misassignment of the ancestral CpG state. A comparison of the results of Subramanian and Kumar (2003) with those of Hellmann et al. (2003) and the (ICGSC, 2005), as well as an analysis of the data used in Keightley et al. (2005b) is presented in Table 2.1. These results clearly show that the rates estimated at intronic non CpG sites are in approximate agreement between these four studies. In addition, in all four studies, the fourfold CpG substitution rate is substantially lower than that estimated at intronic CpG sites, as predicted by the simulations presented in this chapter. However, the rate estimated at non CpG fourfold sites in the dataset of Subramanian and Kumar (2003) is substantially higher (40-60%) than that estimated by the other studies. This discrepancy could result from a number of different sources. Firstly, the results of Subramanian and Kumar (2003) may simply result from a degree of sampling error. Assuming 10-fold CpG hypermutability and 1.5% total pairwise divergence, the simulation results in this chapter predict ~15% elevation of the intronic non CpG substitution rate over the fourfold non CpG substitution rate. As such, the sample size of Subramanian and Kumar (2003) may be inadequate to detect this relatively small difference. However, although the dataset of Subramanian and Kumar (2003) is relatively small compared with those others presented in Table 2.1, the estimate of the fourfold non CpG substitution rate is derived from a reasonably large number of aligned sites (12,473). Perhaps a more likely explanation is that at least some of the intronic sites they used are under some purifying selection. Subramanian and Kumar (2003) appear not to exclude any intronic sites from their analysis, although it is known that at least some of these sites (proximate to the 5′ 3′ splice sites, and within intron 1) are selectively constrained

(Keightley and Gaffney 2003;*see* Chapter 4 and 5), sometimes substantially. If this is the case, excluding these selected, would elevate the intronic substitution rate and produce the expected ratio between fourfold and intronic non CpG substitution rates. This explanation is not ideal however, given that it is the fourfold, rather than intronic, non CpG substitution rate which differs most dramatically from that estimated by other studies. Nonetheless, it is almost certain that some intronic sites in the dataset of Subramanian and Kumar (2003) are under selection. In order to resolve this discrepancy, it is necessary to first identify the genes used by these authors, which were unfortunately unavailable at the time of writing. This would reveal any functional or compositional biases that are particular to the dataset. Secondly, it is necessary to locate and exclude those intronic sites that are likely to be under selection (splice-related sites and all sites in intron 1), to reveal their impact on the analysis.

Table 2.1.: The non CpG substitution rates from ICGSC (2005) were estimated **by eye** from Figure 10 of the manuscript. The fourfold CpG rate was estimated from the 0.0092:0.152 ratio of CpG divergence stated in the Supplementary Notes S2 of ICGSC (2005). The intronic CpG rate is stated to be 50% higher than the fourfold CpG rate in the text (pg 76, paragraph 7). Substitution rates estimated for Subramanian and Kumar (2003) were estimated from the original dataset by the present author using the data kindly provided by Dr. Sankar Subramanian.

Reference	No. genes	$K_{4\ nCpG}$	$K_{Int\ nCpG}$	$K_{4\ CpG}$	$K_{Int\ CpG}$
Subramanian and Kumar (2003)	81	0.0088	0.0084	0.095	0.123
Hellmann et al. (2003)	1845	0.0054	0.008	0.082	0.121
ICGSC (2005)	13454	0.0055	0.0079	0.091	0.137
PDK human-chimp dataset	1260	0.0062	0.0088	0.085	0.151

These results also have implications for the estimation of the rate of evolution within CpG dinucleotides. The potential for overestimation of the CpG substitution rate is high between closely related species. This has been attempted using CpG/non CpG assignment by previous studies (e.g. Ebersberger et al., 2002; Hellmann et al., 2003). It is likely that such studies will have overestimated the rate of substitution at CpG sites, again due to misassignment issues. It is clearly preferable to implement a method which explicitly models context-dependent evolution, such as that proposed by Arndt et al. (2003a), Lunter and Hein (2004) or Hwang and Green (2004).

Despite this, evidence of selection from a variety of sources at mammalian synonymous sites is accumulating (Parmley et al., 2006; Chamary and Hurst, 2005). Analysis of real sequence data would support this conclusion (*see*

Chapter 5). However, these estimates are considerably lower than those estimated using CpG/ non CpG assignment non CpG dinucleotides only (26% and 30%, respectively) and also those estimated by the ICGSC (2005) and by Hellmann et al. (2003), as mentioned above.

Finally, whilst this analysis was restricted to fourfold degenerate and intronic noncoding sequence, the effects of CpG and non CpG misassignment across small evolutionary distances are primarily dependent on differential CpG frequencies, and so these results will apply to any comparison of substitution rates where this is the case. These results recommend against the adoption of *ad hoc* methods of ancestral state assignment.

3. The Scale of Mutational Variation in Murids

The work in this Chapter has been published (Gaffney and Keightley, 2005).

3.1. Introduction

Much evidence now suggests that the point mutation rate varies considerably across the mammalian genome. Studies of nucleotide substitution rates have revealed considerably more variation in the substitution rate than expected by chance at synonymous sites (Wolfe et al., 1989; Matassi et al., 1999; Malcom et al., 2003; Chuang and Li, 2004), within long alignments of primate intergenic sequence (Chen et al., 2001; Smith et al., 2002; Silva and Kondrashov, 2002; Ebersberger et al., 2002) and mammalian repetitive sequence (Waterston et al., 2002; Hardison et al., 2003). This is important for a number of reasons. Firstly, in neutrally evolving DNA, the underlying mutation rate determines the expectation that a sequence will be conserved between two or more sister species by chance. Variation in the mutation rate means that the expectation of evolutionary conservation changes from region to region. This is relevant to studies involving "phylogenetic footprinting" to locate putatively functional regions within noncoding DNA. The identification of such regions could be improved if was known *a priori* which regions are expected to be evolving more slowly. In addition patterns of mutational variation provide information on the processes that are likely to be driving mutation rates.

The regional mutation hypothesis proposes that different regions of the vertebrate genome diverge at different rates (Filipski, 1988). Previous studies have provided evidence that mutation rates vary between chromosomes (Wolfe et al., 1989; Lercher et al., 2001; Malcom et al., 2003; Makova et al., 2004). One notable feature is the apparent reduction in the rate of point (McVean and Hurst, 1997; Ebersberger et al., 2002; Waterston et al., 2002) and indel substitution (Makova et al., 2004) on the X chromosome, relative to the autosomes. This reduction may reflect the primarily male origin of most mutations, although the evidence on this point is inconsistent (McVean and Hurst, 1997; Lercher et al., 2001). In addition, there is evidence

that variation in the mutation rate also occurs along the length of a chromosome (Wolfe et al., 1989; Chuang and Li, 2004). Although mutational variation has been studied at at intra- and inter-chromosomal levels, an unresolved problem is the relative importance of chromosome identity and position within a chromosome in determining the underlying mutation rate. Of particular relevance to this question is the scale of "local similarity" of mutation rates. If the domain or 'unit' of mutational variation is considerably smaller than a chromosome and substantial inter-domain variability exists, this would suggest that position within a chromosome is a more important factor in determining the mutation rate. This conclusion is reversed if mutation rates are relatively invariant across the length of a chromosome. One way to address this question is to estimate the distances at which mutation rates are "locally similar" or autocorrelated.

One of the first studies to address the issue of local similarity of evolutionary rates compared estimates of the synonymous divergence (K_S) from human-mouse gene orthologues within 1cM of each other, and concluded that there is evidence for the existence of 'evolutionary rate units' between which substantial variation exists (Matassi et al., 1999). Lercher *et al.* (2001) extended this analysis to a larger dataset and found that significant autocorrelation of K_S extends from 1cM to entire chromosomes in a human-rodent comparison. Although it may seem unexpected that mutation rates would remain approximately constant across entire chromosomes, this situation does appear to exist in yeast (Chin et al., 2005). Such a large scale of autocorrelation would seem to reject a substantial role for within-chromosomal mutational heterogeneity and suggests that the majority of mutational variation occurs between chromosomes. However, more recent work has suggested that synteny blocks (i.e. regions in which gene order has been conserved between species) may represent a more meaningful 'unit' than whole chromosomes (Malcom et al., 2003; Webster et al., 2004). Malcom *et al* (2003) found that although a weak effect of chromosomal identity is evident from both human-mouse and mouse-rat comparisons, this is confounded by substantial within chromosome variation. These authors indicate that differences between synteny blocks on the same chromosome outweigh those observed between chromosomes. Additional support for a subchromosomal mutational scale comes from Chuang and Li (2004) who use a human-mouse comparison to show that local similarity in mutation rates extends to approximately 10Mb. The relevance of a chromosome as an evolutionarily distinct entity is uncertain,

however, particularly between highly diverged species such as human and mouse, for which genome sequencing projects have revealed many large scale rearrangements (Nadeau and Taylor, 1984; Hudson et al., 2001; Waterston et al., 2002).

Many of the above studies have used synonymous substitution rates to examine patterns of mutational variation. However, synonymous sites comprise a small fraction of most mammalian genomes and may misrepresent mutational processes outside of coding sequence. In addition, the importance of sequence context effects, in particular CpG hypermutability, is becoming increasingly apparent (Arndt et al., 2003a,b; Arndt and Hwa, 2005). It has been shown elsewhere (Chapter 2) that, because mammalian fourfold degenerate synonymous sites can be enriched for the CpG dinucleotide, compared to the majority of the rest of the genome, they have a higher mutation rate than the rest of the genome. Furthermore, evidence is accumulating that selection, perhaps related to mRNA splice efficiency or mRNA stability, is operating at some mammalian synonymous sites (Eyre-Walker, 1999; Keightley and Gaffney, 2003; Willie and Majewski, 2004; Chamary and Hurst, 2004; Keightley et al., 2005b; Parmley et al., 2006).

For these reasons it is desirable to investigate mutational variation outside of coding sequence. Some authors have sought to address this by utilising long human-chimpanzee alignments of intergenic sequence (Smith et al., 2002; Ebersberger et al., 2002; Webster et al., 2004). Webster *et al* (2004) estimated the extent of local similarity using substitution rates at ancestral repeat (AR), intronic and intergenic sites from a human-chimp alignment of 14Mb from human chromosome 7. Their results indicate that the most significant local similarity of mutation rates occurs at a scale of 1-2Mb. However, they did not investigate the rate of decay of this local similarity. Furthermore, it is becoming increasingly apparent that some of the noncoding, nonrepetitive portion of the mammalian genome, assumed to be neutral in the above studies, may be under selection (Thomas et al., 2003; Waterston et al., 2002; Bejerano et al., 2004). Smith *et al.* (2002) and Webster *et al.* (2004) argue that such selected regions should have little influence on substitutional variation in closely related species. However, minimally diverged species are more susceptible to the influence of ancient polymorphism in the last common ancestor, and selection in noncoding DNA does become relevant when considering alternative, more distantly related taxa, such as mouse and rat. Thus, in these species pairs, long intergenic

alignments are not ideal for the study of mutational variation. One alternative is to focus on the remnants of repetitive elements that are were inserted in the last common ancestor (e.g. Waterston et al., 2002; Hardison et al., 2003). The use of these ancestral repeats is appealing because, of all classes of noncoding DNA, they are the most likely candidates for neutrality (Ellegren et al., 2003). Additionally, the large quantities of repetitive sequence in mammalian genomes allow for analysis of mutational variation on much finer scales than would be possible using rates of synonymous substitution.

A dataset of repetitive elements present in the last common mouse-rat ancestor was collected. Using these data the following questions were addressed : (i) What is the scale of autocorrelation of rodent mutation rates? (ii) At this scale, what is the ratio of inter- to intra-chromosomal mutation rate variation? Answers to these questions are important to accurately quantify mutational variation and improve our understanding of the processes which may cause point mutation. Furthermore, information on the scale of mutational variation is important in establishing a robust null hypothesis for comparative genomics methods.

3.2. Materials & Methods

Data

Most mammalian transposable elements can be divided into four broad classes: Short INterspersed Elements (SINEs), Long INterspersed Elements (LINEs), Long Terminal Repeat (LTR) retroposons and DNA transposons. All SINE, LINE, LTR and DNA repetitive elements in build 33.1 of the mouse genome were identified using RepeatMasker (http://www.repeatmasker.org/). Those repetitive elements which were inserted prior to the mouse-rat divergence were identified as follows. 250bp of sequence upstream and downstream of the identified mouse repeat was extracted. Any repetitive sequence in these flanking sequences was masked, also using RepeatMasker. In order to ensure that matches were achieved using reasonable lengths of sequence, any element which did not contain at least 50 consecutive bases of unique, nonrepetitive sequence in both its adjacent flanking regions was excluded. Following masking, the remaining unique sequence was compared to the rat chromosome(s) syntenic to the mouse chromosome on which the repeat originated using BLASTN (Altschul et al., 1997). Chromosomal synteny was as defined in Figure 4 of the IRGSC (2004). The following criteria were used to accept or reject BLAST hits of pairs of flanking

sequence. (i) Hits with E-values of greater than 10^{-5} were rejected. (ii) Hits were only accepted if both flanks had a single unique match on the same rat contig. (iii) It was also ensured that flank matches extended to within 50bp of the start or end of the transposable element. Fulfilment of these criteria indicated that the sequence surrounding the mouse repeat in question was present in the last common murid ancestor. The region between the outer limits of the matched flanks was then extracted from the appropriate rat chromosome of NCBI build 3.1 of the rat genome and aligned to the original mouse flanks and repetitive element sequence using AVID (Bray et al., 2003). The presence of a well-aligned sequence in rat opposite the original mouse repeat in the alignment indicated that the transposable element in question was inserted *prior* to the mouse-rat divergence.

Estimation of substitution rates

Nucleotide substitution rates were estimated for each ancestral repeat and its flanking sequence, correcting for multiple hits using the Tamura-Nei method (Tamura and Nei, 1993). Many transposable elements are GC and CpG rich, and this may affect nucleotide substitution rates, depending on the region of insertion of the element. In addition, analysis of the composition and age of large numbers of repetitive elements in the human genome indicated that element GC content tends to decay over evolutionary time (Lander et al., 2001). This effect violates the assumption of ancestral compositional equilibrium, common to the majority of models used to estimate substitution rates. It is likely, however, that for moderately diverged species, such as mouse and rat, relatively little GC content decay will have occurred since the two species split. Of greater concern is the fact that many mammalian repetitive consensus sequences contain hypermutable CpG dinucleotides at a substantially higher frequency than the genome at large. Hypermutability of CpG dinucleotides in vertebrates is well documented and poses a problem for the estimation of substitution rates using ancestral repeats. Following insertion, CpG dinucleotides within elements are by far the most likely sites to mutate. However, ancient elements will have experienced most CpG-related changes prior to mouse-rat divergence, whereas those more recent insertions may appear to be evolving at an inflated rate due their comparatively higher CpG content. This effect could produce covariation between the age of element insertion and overall divergence, with more CpG-rich, recently inserted elements diverging proportionally faster than their older counterparts. Although there have been recent advances in incorporating context dependency into

models of sequence evolution (Arndt et al., 2003a; Siepel and Haussler, 2004), in this study these issues were addressed by estimating nucleotide substitution rates in three alternate ways: using all sites, at those sites not preceded by a C or followed by a G (non CpG-prone sites) and by counting only A↔T and G↔C changes. The latter two categories are likely to be the least affected by CpG context effects (Chapter 2) and compositional change and allowed the impact of these factors on the results to be assessed.

Mean Chromosome Divergence

The mean chromosomal divergence was calculated treating all elements in a chromosome as a single sequence and summing differences and sites across all elements. Estimates were also corrected for multiple hits using the method of Tamura and Nei (1993). In order to estimate confidence intervals for the average chromosomal substitution rate 1000 bootstrap datasets were generated for each chromosome. Because adjacent substitution rates are autocorrelated, each independent observation in the bootstrap datasets was the substitution rate estimated across all elements in a 2Mb block, in order to minimise dependence between observations. The mean chromosomal divergence was calculated for each dataset and the bootstrap distribution of these was used to estimate 95% confidence intervals for each mouse chromosome. Bootstrap datasets were generated using the 'boot' library in R (R Development Core Team, 2004).

Local Similarity

To investigate the scale of local similarity of substitution rates we divided the mouse genome into 5kb and 100kb blocks and estimated an average block substitution rate by taking a weighted (by number of sites) average of the substitution rates of all elements found within a block. The autocorrelation of substitution rates across blocks was estimated. The k^{th} order autocorrelation, ρ_k is given by (Box et al., 1994):

$$\rho_k = \frac{\sum_{i=1}^{N-k}(K_i - \bar{K})(K_{i+k} - \bar{K})}{\sum_{i=1}^{N}(K_i - \bar{K})^2} \tag{3.1}$$

where, in this study, N is the total number of blocks in the genome and K_i is the substitution rate in block i. In order to provide critical values for the sampling distribution of ρ under the null hypothesis of no relationship between

the evolutionary rates of adjacent blocks, ρ was estimated for 1000 datasets in which block order was randomised. Following Matassi et al. (1999) and Lercher et al. (2001) the impact of local GC content on the observed pattern of autocorrelation was assessed using datasets in which blocks were randomised according to their GC content. Because of the nonrandom pattern of insertion of transposable elements, in all cases elements were permuted while maintaining the structure of the original dataset, e.g. any empty blocks in the real data were maintained as empty blocks in all the randomised datasets. Local GC content was estimated as the average GC content of all masked mouse and rat flanking sequences within a block. Blocks were then assigned to one of a number of GC content classes and randomly permuted only with blocks in the same GC content class where each GC content class contained 5% of the dataset.

To investigate the mean 'unit' of mutational variation the partial autocorrelation of substitution rates averaged across 100kb blocks was estimated. Partial autocorrelation between the mean substitution rate in block x_i and block x_{i+k}, where k is the lag, is the autocorrelation that is not explained by the 'propagation' of lower-order lags $(k-1, k-2, ...)$. In this study, partial autocorrelation of nucleotide substitution rates becomes nonsignificant at the point beyond which all observed similarity of substitution rates can be explained by autocorrelation of rates across smaller distances. All partial autocorrelations were estimated in R (R Development Core Team, 2004). Significance of partial autocorrelations was again assessed using 1000 datasets in which block order was randomised. The partial autocorrelation of substitution rates was estimated in both ancestral repeat and flanking sequence up to an interval distance of 5Mb.

Between and within chromosome variation

The male-to-female mutation rate ratio, α, was estimated using the following formula:

$$\alpha = (3R - 4)/(2 - 3R) \tag{3.2}$$

(Miyata et al., 1987), where R = X/A, and X and A are the mean substitution rates at all sites on the X chromosome and across all the autosomes, respectively.

In order to quantify between and within chromosome mutational variation, the data were fitted to a variety of linear models using the nlme library in R

(R Development Core Team, 2004). Substitution rates in ancestral repeats and flanking sequences were grouped by location into blocks of increasing size from 25kb to average chromosome size (125Mb). The significance of regional effects in explaining variation in the substitution rate was tested by comparing two models:
Model 1:

$$y_{ij} = \beta_i + \epsilon_{ij} \tag{3.3}$$

Model 2:

$$y_{ijk} = \beta_i + \beta_i(b_{ij}) + \epsilon_{ijk} \tag{3.4}$$

In model 1, substitution rate, y_{ij}, is described by an effect of chromosome i (β_i) and a random error term (ϵ_{ij}). In model 2, substitution rate, y_{ijk}, is again described by a mean chromosomal rate, and also by an mean 'regional' rate or effect of block j, b_{ij}. This block effect is modelled as a normally distributed random effect, with the bracket notation denoting that block effects are nested within chromosomes, i.e. as a random variable representing the deviation from the chromosomal mean rate. If substantial regional effects exist then model 2 will provide a significantly better fit to the data than model 1. Both models were fitted to the data using restricted maximum likelihood, as this provides unbiased variance and covariance estimates.

Significant chromosomal effects were also tested for by comparing the fit of Model 2 to the data with the following model (Model 3), which includes a term for a random regional effect (b_i) only:
Model 3:

$$y_{ij} = b_i + \epsilon_{ij} \tag{3.5}$$

If chromosome identity significantly affects variation in substitution rate, Model 2 will provide a better fit to the data than Model 3. Model 2 and Model 3 were fitted to data both including and excluding the X chromosome, which is a chromosomal outlier. In this case the data were fitted using 'full' maximum likelihood as Model 2 and Model 3 differ in their fixed effects specification and their log-restricted-likelihoods cannot be compared (Pinheiro and Bates, 2000).

For all comparisons the Akaike Information Criterion (AIC) was used to assess the fit the model to the data. The AIC is a model selection criterion which incorporates information about the fit of the model to the data and the model

complexity:

$$AIC = -(-2l(\hat{\theta}|\boldsymbol{y}) + 2n_{par}) \tag{3.6}$$

where $l(\hat{\theta}|\boldsymbol{y})$ is the log-likelihood of the model, $\hat{\theta}$, given the data, $\boldsymbol{y}$ and n_{par} is the number of parameters in the model (Pinheiro and Bates, 2000).

Simulations

In order to assess the efficacy of a linear model in estimating the regional scale and magnitude of mutational variation a simple simulation protocol was employed. Chromosome effects were simulated as random draws from a normal distribution, of mean = 0.1596 (the overall mean divergence estimated from ancestral repeats) and standard deviation = 0.0048 (the between chromosome standard deviation, estimated from fitting model 1 to the ancestral repeat data). In the null model (no regional effects) element substitution rates were simulated as random draws from a normal distribution of mean equal to the mean of the chromosome on which they were situated, and standard deviation = 0.0501 (residual from model 1). In the second model, 'block' regional effects were simulated as random draws from a normal distribution of mean equal to the mean of the chromosome on which they were located, and standard deviation = 0.0150 (the between block standard deviation estimated by fitting model 2, including a term for 1Mb blocks, to the ancestral repeat data). Within a 'block' element substitution rates were simulated as random draws from a normal distribution of mean equal to the simulated block substitution rate, and standard deviation = 0.0487 (the residual from fitting model 2, including a term for 1Mb blocks, to the ancestral repeat data). Blocks of 100kb, 1Mb and 5Mb in size were simulated. Repeat insertions were simulated as a Poisson process across a chromosome. These simulated data were subsequently analysed as described in Section 3.2.

3.3. Results

A total of 55 Mb of repetitive sequence was extracted and aligned. This can be broken down into the following contributions from various classes of repetitive element: 17.5 Mb of SINE, 13.0 Mb of LINE, 21.0 Mb of LTR and 3.7 Mb of DNA transposon. The proportions of aligned sequence derived from each repeat family appears approximately consistent across autosomes (Figure 3.1). However, LINE

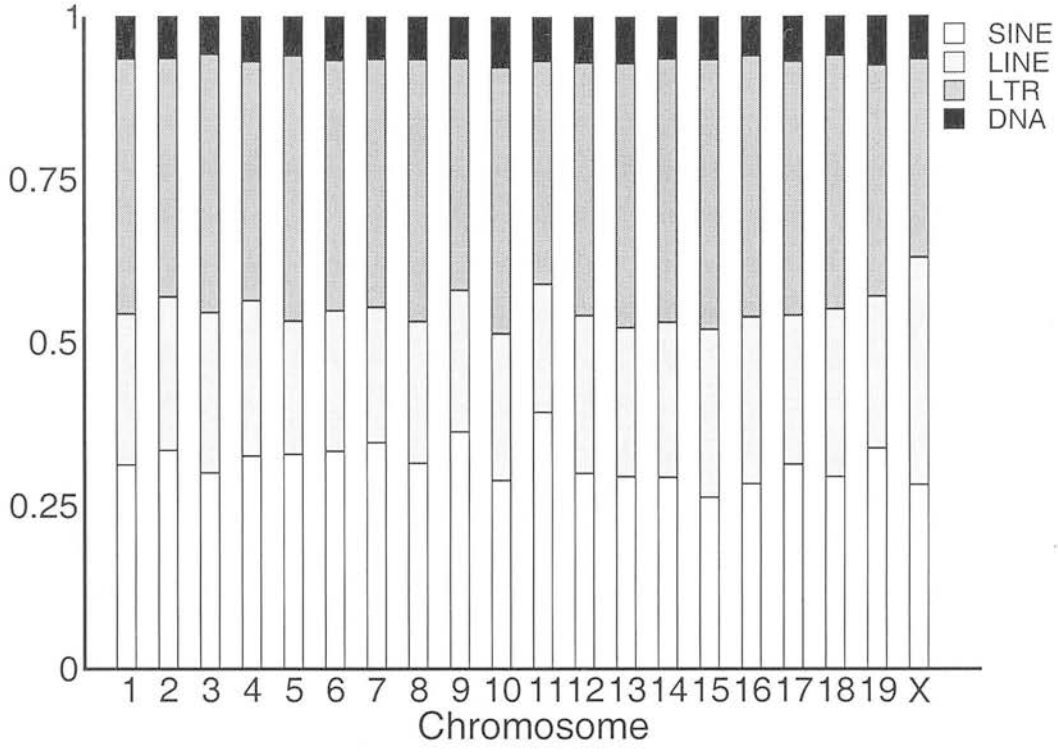

Figure 3.1.: Figure 3.1: Proportion of total sequence per mouse chromosome contributed by each repeat class.

elements appear to be significantly more prevalent on the X chromosome (44% more frequent on the X chromosome than on the mean autosome; $P < 0.0001$) than the autosomes. This would suggest either that LINE elements have been more active on the X chromosome or the rate of deletion of LINEs was less than on the autosomes in the common ancestor of mouse and rat. There is some evidence to suggest that the former scenario is more likely, as it seems that some retrotransposing sequences preferentially target the X chromosome (Khil et al., 2005). It may also be that LINEs play a role in X chromosome inactivation (Waterston et al., 2002; Bailey et al., 2000).

Between chromosome variation

The average chromosomal divergence was estimated at all sites and at sites not preceded by a C or followed by a G (non CpG-prone sites) for each mouse chromosome (Figure 3.2). Non CpG-prone sites are the least likely to have been part of a hypermutable CpG dinucleotide (*see* Chapter 2), and therefore the least affected by potential covariation between nucleotide divergence and age of transposable element insertion. The X chromosome appears to be evolving more slowly at all sites than any of the autosomes and the estimated male-to-

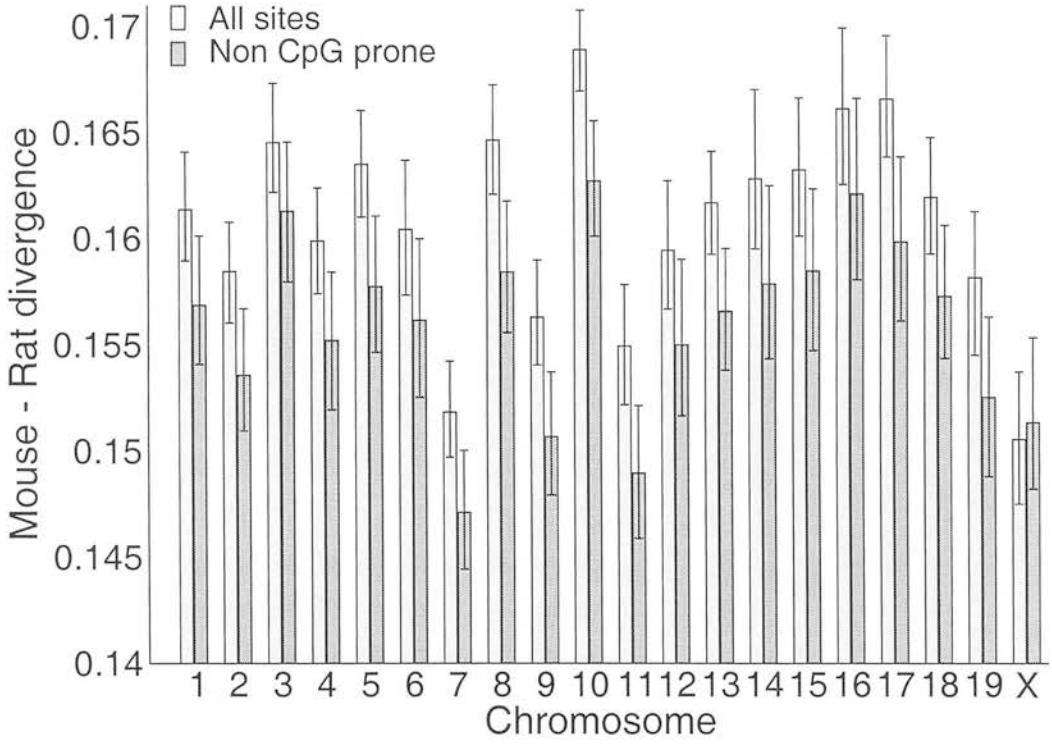

Figure 3.2.: Figure 3.2: Estimated average nucleotide substitution rates at all sites and non CpG-prone sites for each mouse chromosome. Bars show the 95% critical values for the sampling distribution of ρ under the null hypothesis of no local similarity of nucleotide substitution rates.

female mutation rate, α, is $\sim$1.5. This is slightly lower than previous estimates in rodents (1.8, Chang et al. (1994);1.9, IRGSC (2004)), but is consistent with the overall picture of weak male-driven evolution in rodents. In comparison, estimated male-to-female mutation rates in longer-lived taxa, such as hominids, are substantially higher (3-6, (ICGSC, 2005)). Rates at non CpG-prone sites are consistently lower than those estimated at all sites for all autosomes. This would suggest that rates inferred at all sites are affected by the elevated mutation rates at CpG dinucleotides and the selection of non CpG-prone sites goes some way to removing this effect. Interestingly, however, this situation is reversed on the X chromosome, where substitution rates at non CpG-prone sites are in fact marginally, although not significantly, higher than those estimated at all sites. This result appears to be roughly consistent within repeat families (Table 3.1).

Table 3.1.: Nucleotide substitution rates at all and non CpG prone sites in the autosomes and X chromosome, by repeat family

	X chromosome		Autosomes	
Site	All	Non CpG-prone	All	Non CpG-prone
Repeat				
SINE	0.154	0.163	0.160	0.162
LINE	0.146	0.148	0.157	0.154
LTR	0.150	0.155	0.162	0.160
DNA	0.135	0.135	0.151	0.145

Scale of local similarity

The scale of local similarity of mutation rates was estimated using the autocorrelation of average substitution rate across a variety of block sizes. Figure 3.3 shows the autocorrelation of nucleotide substitution rates at all sites between blocks of 5kb and 100kb extending over intervals from 10kb to 1Mb and 200kb to 20Mb, respectively. Autocorrelation of rates across 5kb blocks (Figure 3.3, a) remains highly significant compared to randomly permuted data across a distance of 1Mb. There is minimal change in autocorrelation from 10kb to 100kb suggesting that little variation in underlying mutation rate exists below 100kb. The low magnitude of autocorrelation across 5kb blocks reflects the relatively noisy estimates of substitution rate obtained from the small number of ancestral repeat sites (295bp on average) within each block. In contrast, the number of sites within the average 100kb block is approximately one order of magnitude larger than that in 5kb blocks (2.3kb on average), so the estimate of the substitution rate is less noisy and the magnitude of autocorrelation higher.

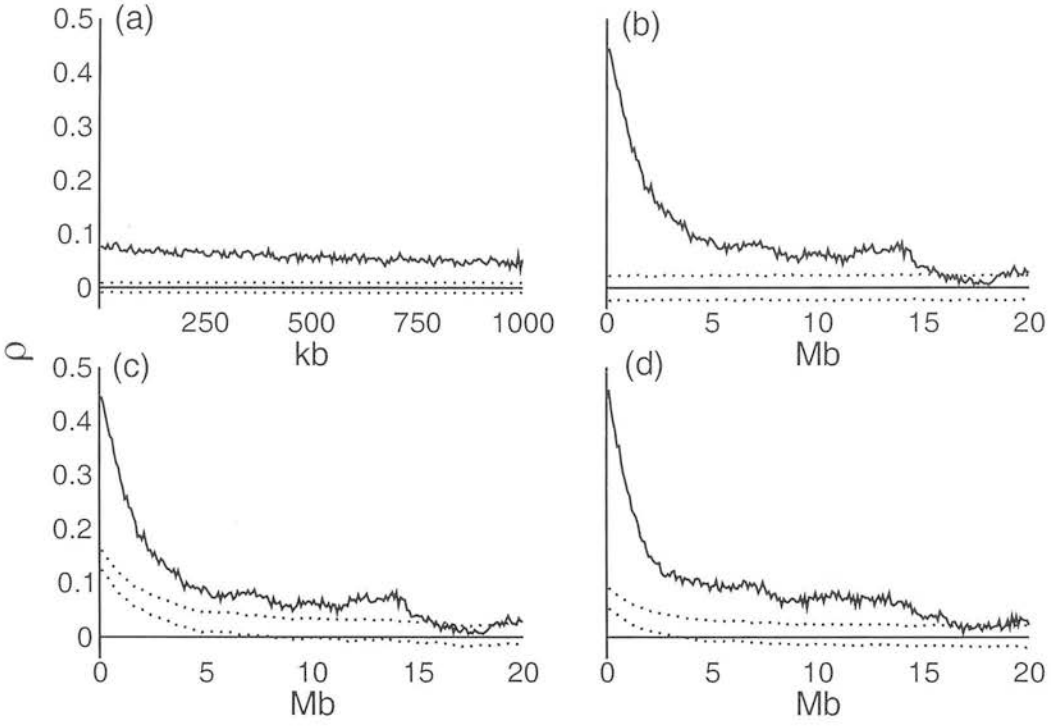

Figure 3.3.: Autocorrelation of nucleotide substitution rates in ancestral repeats (a-c) and ancestral repeat flanking sequence (d) across 5kb (a) and 100kb (b-d) blocks. Substitution rates were estimated at all sites. Dotted lines show the upper and lower bounds of the 95% confidence interval of autocorrelation under the null hypothesis of no dependence of rates between blocks. Blocks were permuted randomly (a, b) and within common GC content intervals (c-d).

Here, there is a slow decay of autocorrelation of substitution rates extending to a distance of 10-15Mb (Figure 3.3, b). It is important to note that autocorrelation in Figures 3.3a and 3.3b show the same proportional change over the same distance. For example, autocorrelation across 5kb blocks decays from ~0.078 to ~0.052 (a decrease of approximately one third) over a distance of 1 Mb; autocorrelation across 100kb blocks decays from ~0.445 to ~0.290 (again a decrease of approximately one third) over the same distance.

The similarity of evolutionary rates between blocks within an interval of 0-15Mb seems to be explained, in part, by the corresponding similarity of average GC content of adjacent blocks. GC content of a block is defined as the mean GC content of all repeat-flanking sequences, in which all repetitive sequence has been masked, in a block. It can be seen from Figure 3.3 c and d that randomly permuting blocks within GC classes still produces a moderately positive autocorrelation in the absence of local structure. This would suggest that local GC content, or one or more covariates of local GC content, influences

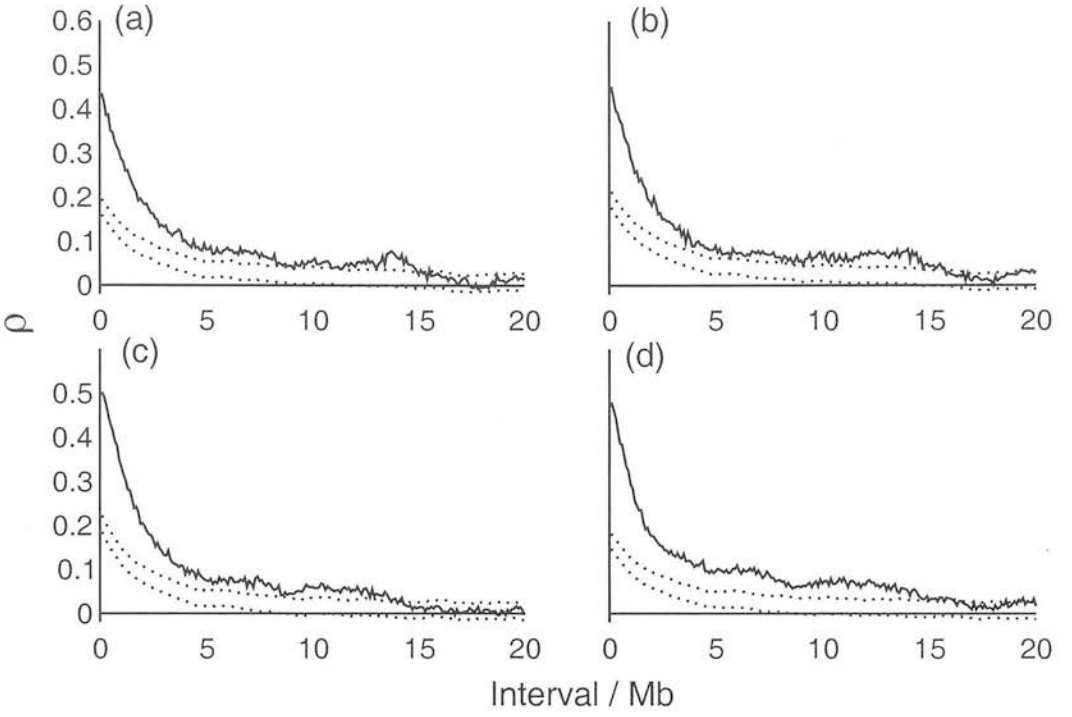

Figure 3.4.: Autocorrelation of nucleotide substitution rates counting only A↔T and G↔C changes (a,c) and at non CpG-prone sites (b,d) in ancestral repeats (a,b) and flanking sequence (c,d) across 100kb bins. Dotted lines show the upper and lower bounds of the 95% confidence interval of autocorrelation under the null hypothesis of no dependence of rates between blocks. Blocks were permuted within common GC content intervals.

neutral substitution rates in both repetitive and non repetitive DNA. However, this similarity does not seem to be as a result of CpG hypermutability or compositional change, since the results were qualitatively similar when rates were estimated at non CpG-prone sites or by counting A↔T and G↔C changes only (Figure 3.4). Given that biased gene conversion appears to primarily affect substitutions that change base composition (Meunier and Duret, 2004) A↔T and G↔C substitutions will be unaffected by this mechanism. Thus this result suggests that variation in the level of biased gene conversion is not responsible for the autocorrelation of substitution rates observed in this study.

The partial autocorrelation of nucleotide substitution rates in both ancestral repeats and flanking sequence was also estimated, averaged across 100kb blocks (Figure 3.5). Plots of partial autocorrelation coefficients show that all autocorrelation over distances greater than 1Mb can be explained by "lower order" autocorrelations below 1Mb. This suggests that the average 'unit' of mutational variation is no larger than ∼1Mb in size. The results are similar in

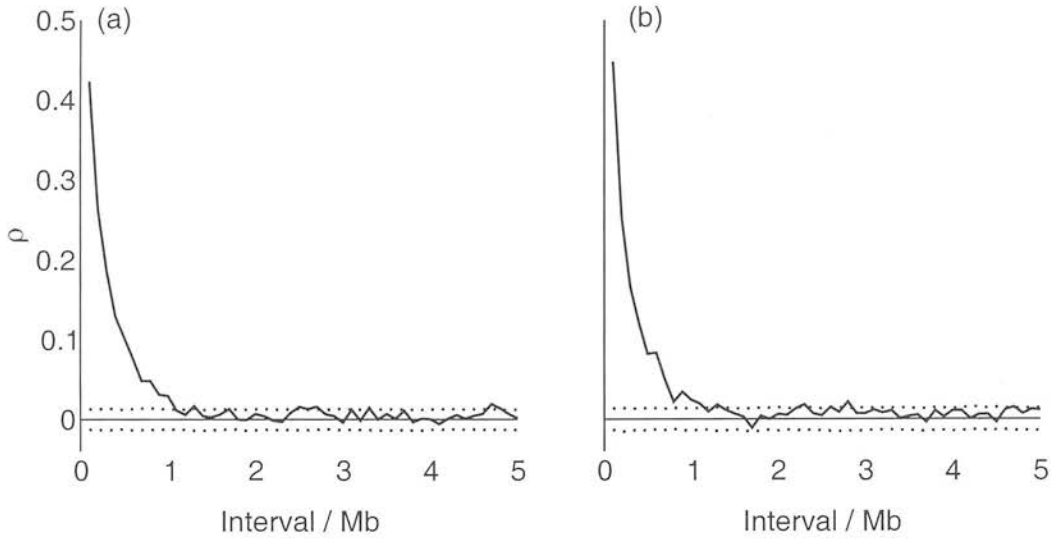

Figure 3.5.: Partial autocorrelation of nucleotide substitution rates in ancestral repeats (a) and flanking sequences (b). Substitution rates are estimated all sites. Dotted lines show the upper and lower bounds of the 95% confidence interval of partial autocorrelation under the null hypothesis of no dependence of rates between blocks.

both repetitive and nonrepetitive sequence (Figure 3.5 a and b, respectively).

Model Efficacy

The efficacy of the model was estimated using a simple simulation protocol. Null model simulation data (no regional effects) were analysed with both Model 1 and Model 2, including a variety of block sizes in the latter (Figure 3.6). Data with simulated regional effects of various physical sizes was then analysed using Model 2 (Figure 3.7). These results are averaged across 250 simulated replicates. Results of this analysis indicate that when regional effects are absent, Model 1 (fixed chromosome effects only) explains the data more parsimoniously than Model 2 (fixed chromosome and random block effects), independent of the block size included in Model 2 (Figure 3.6).

When regional effects of varying sizes are simulated, Model 2 provides a substantially better fit to the data (Figure 3.7), as is the case with in the real data. The best fitting mixed effects model (i.e. the model with the lowest AIC) is that which includes a block size closest to the true simulated block size. The between block variance estimated from the best fitting mixed effects model also appears to be a reliable estimator of the simulated value (Figure 3.8).

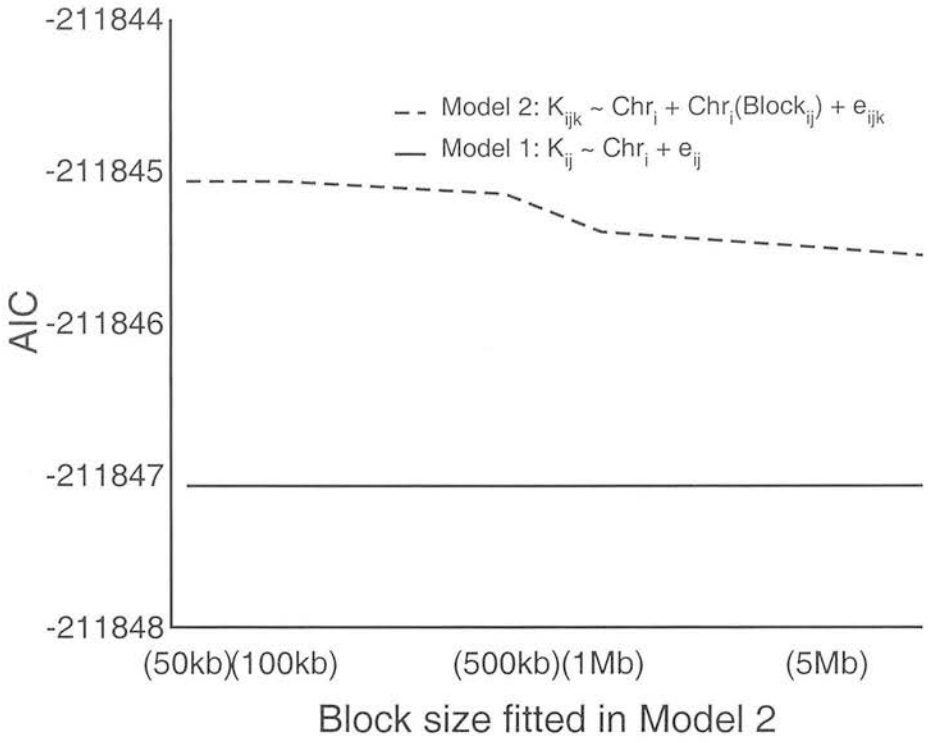

Figure 3.6.: The AIC returned by the two linear models, model 1 fitting just chromosome fixed effects, model 2 fitting fixed chromosome and random regional effects to data with simulated chromosomal but no regional effects. AICs returned by model 2 are shown fitting a variety of different block sizes on the X axis.

Within and between chromosome mutational variation

The data were initially fitted to two linear models, one including terms for fixed chromosomal and random regional effects, and the other including a chromosomal effect only. The magnitude of within chromosome mutational variation was estimated as the variation between levels of the random regional effect in the former model. Model fit was assessed using Akaike's Information Criterion (AIC).

It appears that all models that included 'regional variation' effects provide a substantially better fit to the data than those including chromosome means alone (Figure 3.9, 3.10, 3.11). This is clearly seen from the decrease in AIC (models with a better fit have a lower AIC) for models including a random regional effect. The AIC for Model 1 fitted to substitution rates estimated at all sites was -844146.7 and -933598.5 for ancestral repeats and flanking sequence data, respectively. Including blocks of 1Mb as a random effect in the Model 2 fitted to the same data, for example, decreases the AIC to -854057.2 for the ancestral repeat data (Figure 3.9) and to -945433.5 for the flanking sequence data.

This is evidence that significant regional variation in neutral mutation rate

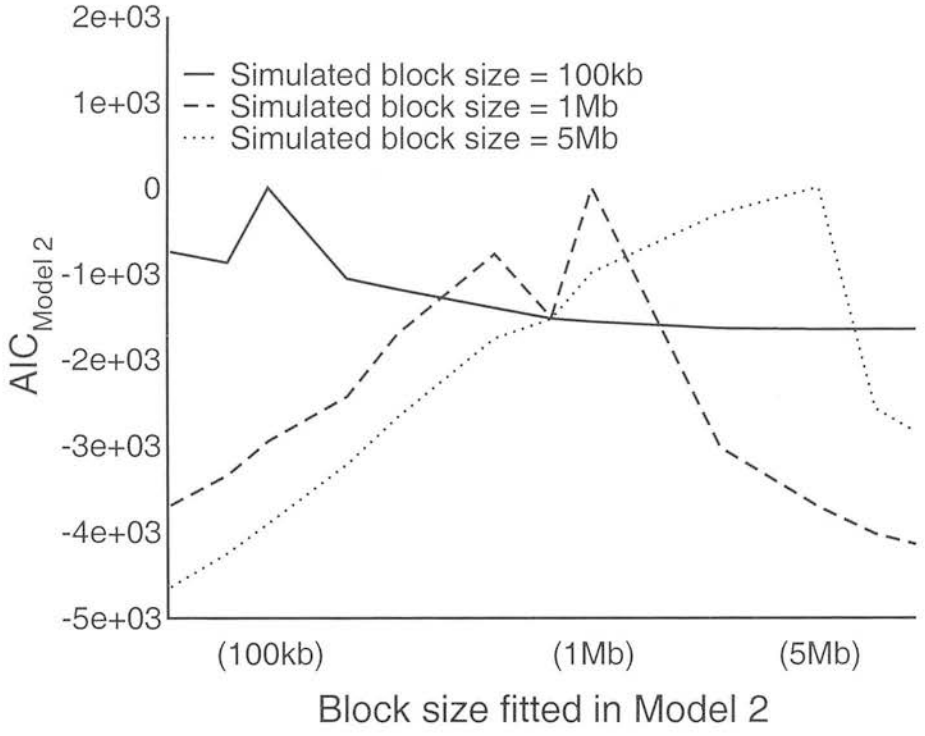

Figure 3.7.: AIC returned fitting model 2, including terms for fixed chromosomal effects and random regional effects of a variety of sizes, to data with simulated chromosomal and regional effects. All AIC values have been normalised to 0 for clarity by subtracting the minimum (i.e. best-fitting) AIC from all data points. Regional effects were simulated in blocks of 100kb, 1Mb and 5Mb in size. The mean AIC (averaged across 250 replicates) returned fitting a model including only fixed effects were -209959.9, -209997.6 and -210215.7 for simulated block sizes of 100kb, 1Mb and 5Mb respectively. In contrast, the lowest mean AIC returned by models fitting regional effects was -210210.2, -210700.4 and -211354.5 for simulated block sizes of 100kb, 1Mb and 5Mb respectively.

does indeed occur along the length of a chromosome. The most parsimonious model (as measured by the AIC) in the analysis includes a block size of 1Mb as a random effect. At this scale the variation between blocks is approximately one order of magnitude greater than that observed between chromosomes (Figures 3.9, 3.10 and 3.11). The between chromosome variance in substitution rate at all sites was 2.28×10^{-5} and 7.71×10^{-7} for ancestral repeats and flanking sequence, respectively, whereas the between block variance in the most parsimonious model is 2.06×10^{-4} in ancestral repeats and 9.53×10^{-5} in flanking sequence. Whilst the substitution rates in ancestral repeats appear more variable than flanking nonrepetitive sequence, the inter- versus intra-chromosomal mutational variation is striking in both categories of site. The results are qualitatively consistent whether rates at non CpG-prone sites are considered or by counting only A$\leftrightarrow$T

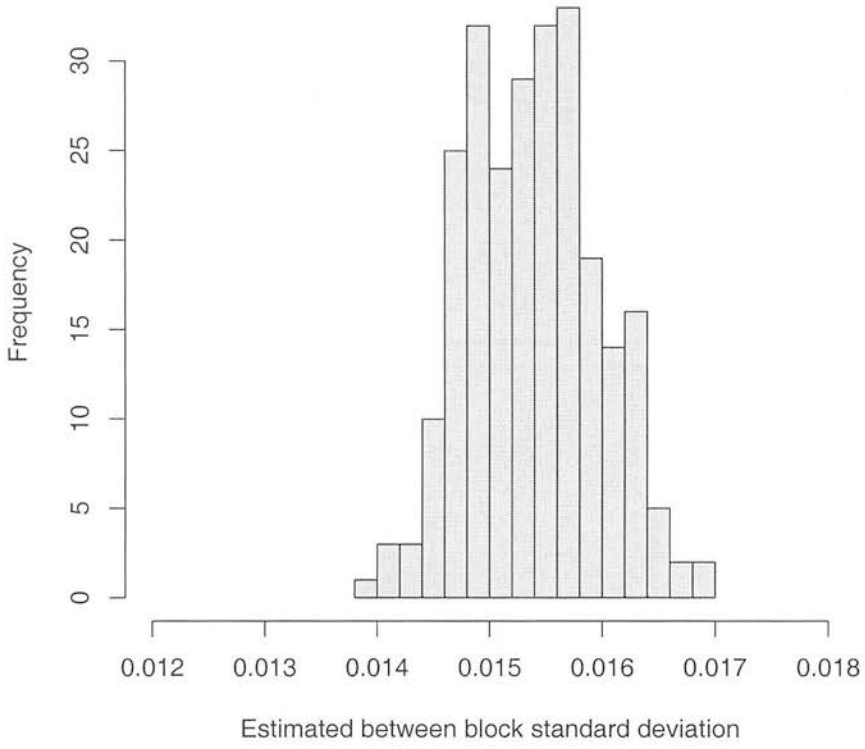

Figure 3.8.: Frequency distribution of the between block standard deviation estimated fitting Model 2 (including a term for a 1Mb regional effect) to simulated data. Estimates are for 250 simulated replicates and the simulated value of the between block standard deviation is 0.015.

and G↔C changes (Figures 3.10 and 3.11 respectively).

It was also determined whether there were significant chromosome effects by comparing the mixed model (Model 2) with a model that includes a term for random regional effects only (Model 3). Regional effects of 1Mb were included in both models. Four different datasets were analysed, consisting of nucleotide substitution rates in ancestral repeats and flanking sequence, including and excluding the X chromosome. The results indicate that model 2 describes the data most parsimoniously in all cases (Table 3.2). The difference in AIC between Model 2 and Model 3 is far smaller, however, (approximately two orders of magnitude) than that observed between Model 1 and Model 2. Differences in AIC between Model 2 and Model 3 drop still further when the X chromosome is excluded. All of these analyses support the conclusion that whilst there exist

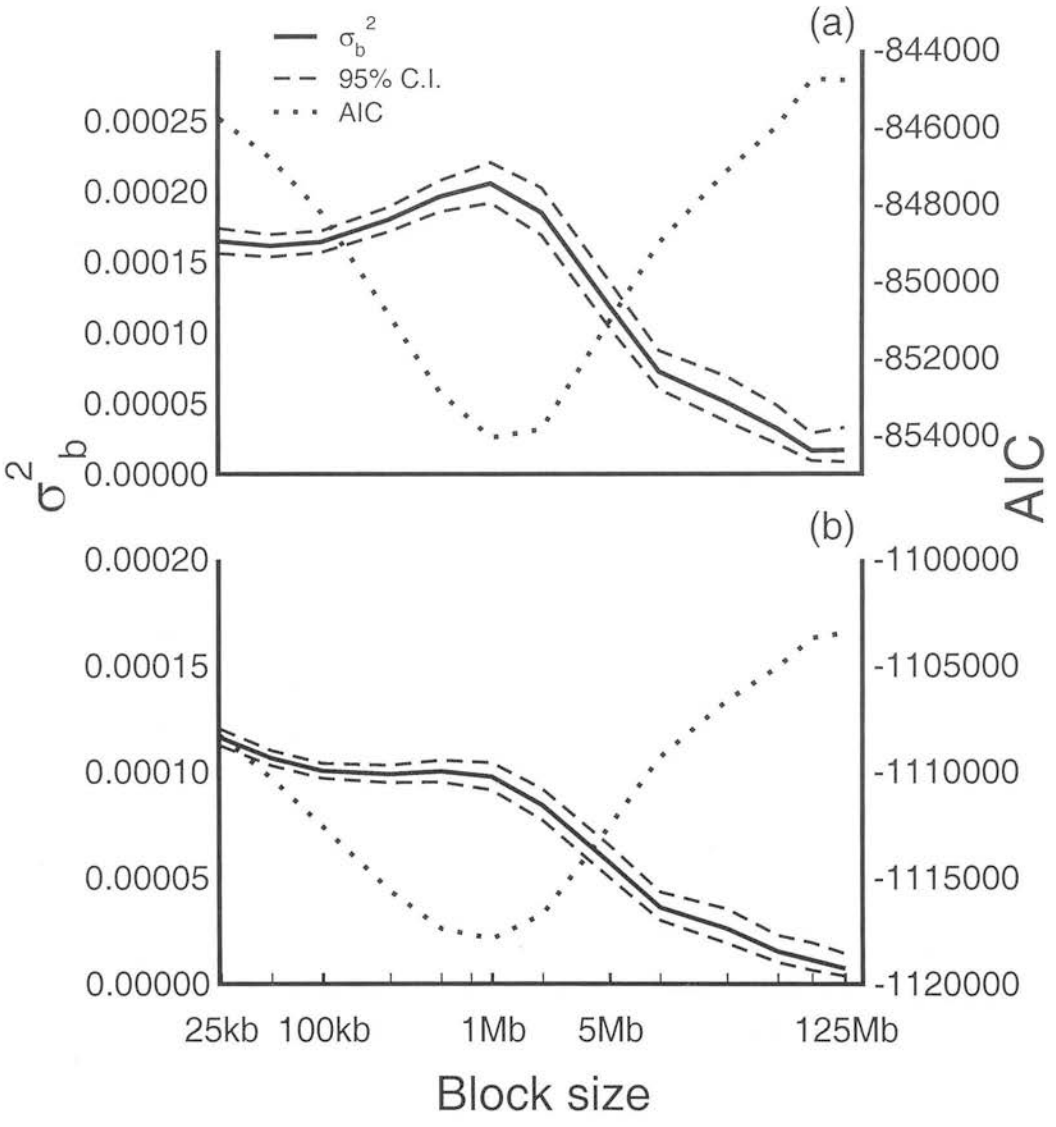

Figure 3.9.: Between block variation (σ_b^2) in substitution rates within ancestral repeats (a) and flanking sequence (b). Substitution rates are estimated at all sites. Between block variances are estimated fitting chromosome as a fixed effect and block as a random effect across different block sizes, from 25kb to 125Mb. Fitted block sizes are plotted on a $\log_{10}$ scale. 95% confidence intervals of the between block variance were as estimated by the lme routine of the nlme package in R. The Akaike Information Criterion (AIC) is shown for each fitted model.

small but detectable chromosomal effects on nucleotide substitution rates, they are far outweighed by intra-chromosomal regional variation.

It should be noted that the mixed model does not explain a large proportion of the variance in substitution rate (~6%) when fitted to data consisting of observations on individual ancestral repeats, as is presented above. However,

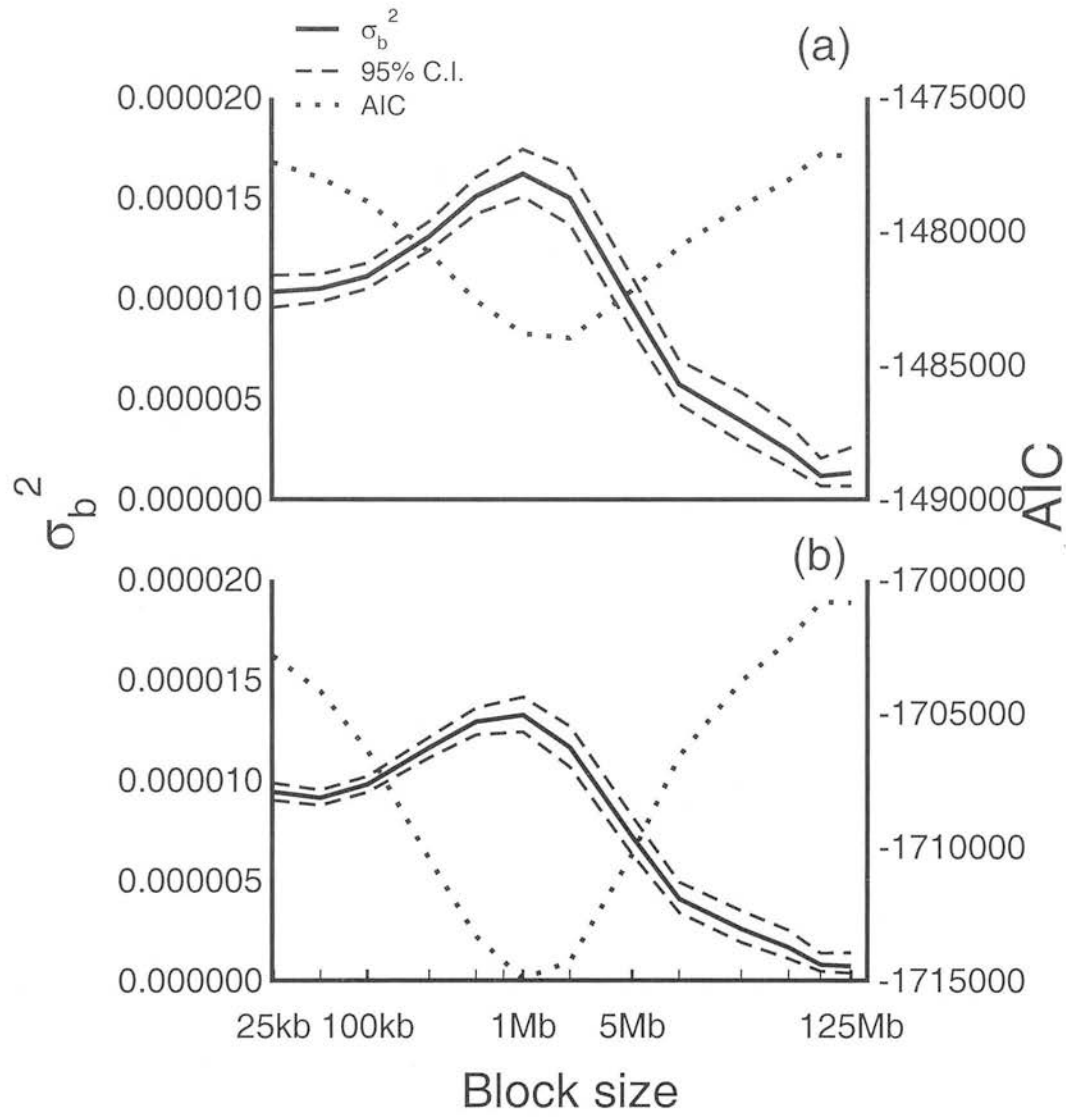

Figure 3.10.: Between block variation (σ_b^2) in substitution rates within ancestral repeats (a) and flanking sequence (b). Substitution rates are estimated counting only A↔T and G↔C changes. Between block variances are estimated fitting chromosome as a fixed effect and block as a random effect across different block sizes, from 50kb to 125Mb. Fitted block sizes are plotted on a $\log_{10}$ scale. 95% confidence intervals of the between block variance were as estimated by the lme routine of the nlme package in R. Akaike's Information Criterion (AIC) is shown for each fitted model. The between chromosome variation estimated counting only A↔T and G↔C changes is 8.2×10^{-7} in ancestral repeats and 4.97×10^{-7} in flanking sequence.

it is likely that much of the residual variation is due to the considerable error involved in inferring substitution rates from such small sequences (on average ~200bp). This is supported by the observation that the proportion of variance explained by the mixed model when fitted to the slightly longer flanking sequences (on average ~362bp) is higher (~9%). If it is assumed that there is minimal

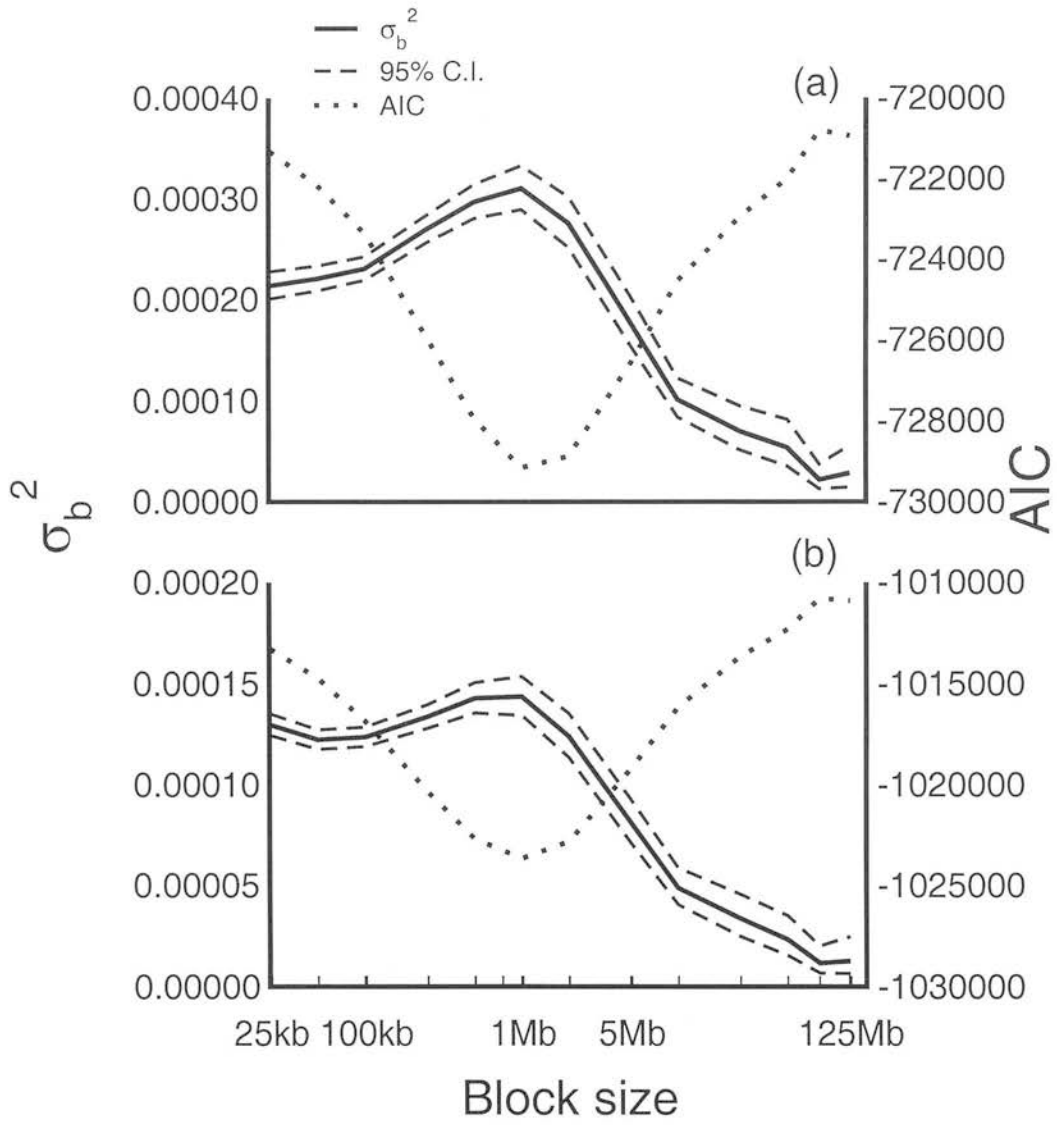

Figure 3.11.: Between block variation (σ_b^2) in substitution rates within ancestral repeats (a) and flanking sequence (b). Substitution rates are estimated at non CpG-prone sites. Between block variances are estimated fitting chromosome as a fixed effect and block as a random effect across different block sizes, from 50kb to 125Mb. Fitted block sizes are plotted on a $\log_{10}$ scale. 95% confidence intervals of the between block variance were as estimated by the lme routine of the nlme package in R. Akaike's Information Criterion (AIC) is shown for each fitted model. The between chromosome variation estimated at non CpG-prone sites is 1.66×10^{-5} in ancestral repeats and 9.07×10^{-6} in flanking sequence.

mutational variation below 50kb and thus treat all ancestral repeats within a 50kb window as a single sequence having a single mutation rate, the mixed model, including a term for a 1Mb regional effect, explains ~25% of the total variation in estimated substitution rate. Thus it is likely that 25% is a reasonable estimate of the proportion of true mutational variation explained by the most parsimonious

Table 3.2.: Akaike Information Criteria for model 2 (chromosomal and regional effects) and model 3 (regional only) when fitted to each of four datasets: nucleotide substitution rates in ancestral repeats and flanking, nonrepetitive sequence, including and excluding the X chromosome. Both models included a term for a 1Mb regional effect.

	All Chromosomes		Autosomes Only	
	Ancestral Repeat	Flank	Ancestral Repeat	Flank
Model 2	-854283.5	-1118129	-803472	-1052114
Model 3	-854133	-1117967	-803349.8	-1052005

model in these analyses.

3.4. Discussion

This study provides further evidence for, and clarification of, the regional mutation hypothesis. The results illustrate that the primary scale over which mutation rates vary is subchromosomal and that intra-chromosomal effects are at least as important as male germline effects as a source of mutational variability, although the latter have received substantially more attention in the literature. The evidence for this conclusion is threefold. Firstly, partial autocorrelations suggest that all long-range (>1Mb) similarity of mutation rates can be explained by 'propagation' of similarity of mutation rates across distances of <1Mb. Secondly, results of the mixed model analysis indicate that within-chromosome mutational variation greatly exceeds variation among chromosomes. Given that chromosomal location of X-linked sequence appears highly conserved between mouse and rat (IRGSC, 2004), it is unlikely that the intra chromosome variation observed could be the result of differences in time spent within the male germline. Thirdly, comparison of models 2 and 3 demonstrates that the effects of chromosome on mean nucleotide substitution rate are small.

The results presented here do not show significant autocorrelation of substitution rates across scales as large as an entire chromosome in murids, as a previous human-mouse study has indicated (Lercher et al., 2001). A possible explanation for this discrepancy is that the mutation pattern has undergone a substantial shift in the lineage leading from the murid common ancestor to human although it is unclear how such an event might have occurred. A more likely possibility is that the greater divergence between human and mouse simply affords greater power to detect such small effects. Notwithstanding, a

recent large scale study of the synonymous substitution rates at approximately 15,000 human-mouse gene orthologues supports the conclusion of local similarity extending to 10-15Mb intervals (Chuang and Li, 2004).

Despite the findings of studies such as Lercher et al. (2001) and Chuang and Li (2004), it is not universally accepted that autocorrelation/local similarity of nucleotide substitution rates exists. In a study of nucleotide substitution rates across a wide range of mammalian genes Kumar and Subramanian (2002) suggest that little local similarity in substitution rates exists. Specifically, they find that the mean difference in fourfold substitution rates between nearby ($\leq$0.5Mb) genes is approximately the same as that between genes located a considerable distance from one another. They argue that reason for the departure of their results from previous findings in similar datasets is the inclusion of genes which have been evolving heterogeneously in different lineages. They argue that the exclusion of such genes, based on a disparity index statistic (Kumar and Gadagkar, 2001), removes "false" local similarity of substitution rates which they suggest is an artifact of heterogeneous evolution. Although Kumar and Subramanian (2002) base their conclusions on local similarity solely upon a dataset of human-mouse gene orthologues, as opposed to mouse-rat orthologues, their results nonetheless require discussion in the context of the results presented in this chapter. One potential reason why Kumar and Subramanian (2002) may have failed to find a significant signal of local similarity is that the statistic they use to identify local similarity (the mean difference in fourfold substitution rates averaged across genes all genes located a certain distance from one another) is unlikely to be powerful enough to detect a weak signal of local similarity. A number of other studies which employ a variety of more powerful methods all find highly significant local similarity both between murids and mouse and human (Lercher et al., 2001, 2004; Chuang and Li, 2004). In particular, Lercher et al. (2004) specifically address the findings of Kumar and Subramanian (2002) in their study of 5212 human-mouse and 4442 mouse-rat gene orthologues. To investigate whether local similarity of nucleotide substitution rates in mammals is merely an artifact of heterogeneous evolution between some gene orthologues, they exclude all genes in their dataset which fail the disparity test proposed by Kumar and Gadagkar (2001). However, in this reduced dataset of homogeneously evolving gene orthologues (Lercher et al., 2004) still find that putatively neutral substitution rates in both linked genes and introns within the same gene are significantly more similar than would expected by chance.

This result directly contradicts the findings of Kumar and Subramanian (2002). Significantly, Lercher et al. (2004) demonstrate that the method employed by Kumar and Subramanian (2002) is not powerful enough to detect the signal of local similarity they find in their own dataset. They suggest that this lack of power is primarily due to the restriction upon comparison of nearest neighbours.

It is interesting to note that whilst the estimates of between chromosome variation estimated in this study ($\sim 2.3 \times 10^{-}5$) are consistent with previous estimates from murid ancestral repeats (e.g. $\sim 3 \times 10^{-5}$; Makova et al. 2004), they are lower than the between chromosome variation recently estimated at synonymous sites (2.7×10^{-4}; Malcom et al. 2003). However, the variance of the estimated mean chromosomal K_S from Malcom et al. (2003) is also somewhat larger than the variance of chromosomal substitution rates estimated from ancestral repeats in this work ($\sim$0.0069 vs $\sim$0.0025). It seems, therefore, that substitution rates at synonymous sites are considerably more variable than rates within ancestral repeat sequences. This may be as a result of selection on some synonymous sites, or interaction between the effects of strong selection on sites adjacent to synonymous sites and context-dependent mutational processes. It is likely that intra chromosome mutational variation would exceed inter chromosome variation if rates were estimated at synonymous sites.

These results raise questions about the biological mechanisms that give rise to new mutations. Mistakes by DNA polymerase in strand replication are thought to be responsible for the apparent mutagenic properties of cell division. The pattern of variation observed here could be explained by two, non mutually exclusive, processes. Firstly, the accuracy of DNA replication may vary regionally along the length of chromosomes. This could elevate or diminish the mutation rate in different regions of the same chromosome. It not obvious, however, whether a specific biological mechanism which could produce regionally varying replication accuracy exists. It is known the quantities of newly synthesised deoxyribonucleotides fluctuates during the cell cycle and it has been proposed that these fluctuations during DNA replication could influence regional variation in the mutation rate, given that different chromosomal regions are replicated at different times (Wolfe, 1991). One straightforward method to test this hypothesis would be to determine the extent of correspondence (if any) between chromosomal replication origins and the nucleotide substitution rate in ancestral repeats.

Clearly, however, if cell division were the major source of most point mutations a substantially larger difference between autosomes and the X chromosome than between regions of the same chromosome would be expected. Secondly other factors, such as structural alterations and spontaneous degradation of nucleotide bases which are unaffected by DNA replication could contribute substantially to the production of single base pair mutations. Such alterations could include processes such as the deamination of methylcytosine to thymine or oxidative base damage caused by oxygen free radicals. That the pattern of variation remains the same when considering substitution rates at non CpG-prone sites (Figure 3.11) would suggest that CpG-derived mutation is not responsible for much of the regional variation observed. It is unclear whether those mutations produced by oxidative base damage can be distinguished by mutations derived from other sources, however. A further factor which may influence variation in underlying mutation rate is the changing nature of chromatin fibre structure across the mammalian genome. One recent study has shown that, contrary to what might be expected, there is an excess of DNA damage in the gene rich nuclear interior, compared with the exterior (Gazave et al., 2005). It is possible, that much of the variation in nucleotide substitution rate observed in this chapter relates to such large-scale chromosomal organisation.

The magnitude of within chromosomal mutational variation highlights the importance of accounting for regionally varying mutation rates in the identification of putatively functional regions of noncoding DNA. Although the coefficient of regional variation in nucleotide substitution rates observed is not large (8.75%, estimated fitting 1Mb regional effects to the data), this still impacts on the null expectation of conservation of a sequence between two species. As an example, assuming that mouse-rat divergence is normally distributed with mean=0.16 and standard deviation=0.014, 95% of divergence scores will be in the range 0.132-0.188. The probability of $\geq$95% sequence identity of a 100bp sequence between two species at the lower 95% bound is over two orders of magnitude larger than the probability of identity for the same sequence at the upper 95% bound. This observation emphasises the importance of estimating neutral mutation rates locally. Additionally, these results illustrate that there is likely to be an effect of sampling when estimating average chromosomal substitution rate solely from genic regions. The majority of mammalian genes reside in GC rich regions (Mouchiroud et al., 1991; Lander et al., 2001) so

even sampling all genes from a chromosome may return a regionally biased estimate of chromosomal evolutionary rate, and any subsamples thereof will potentially exaggerate this bias. Clearly, in order to accurately estimate an average chromosomal mutation rates one must sample from all regions of a chromosome, not just genic regions. Intra-chromosomal mutational variation could explain some disparities between previous estimates of average X and autosomal substitution rates.

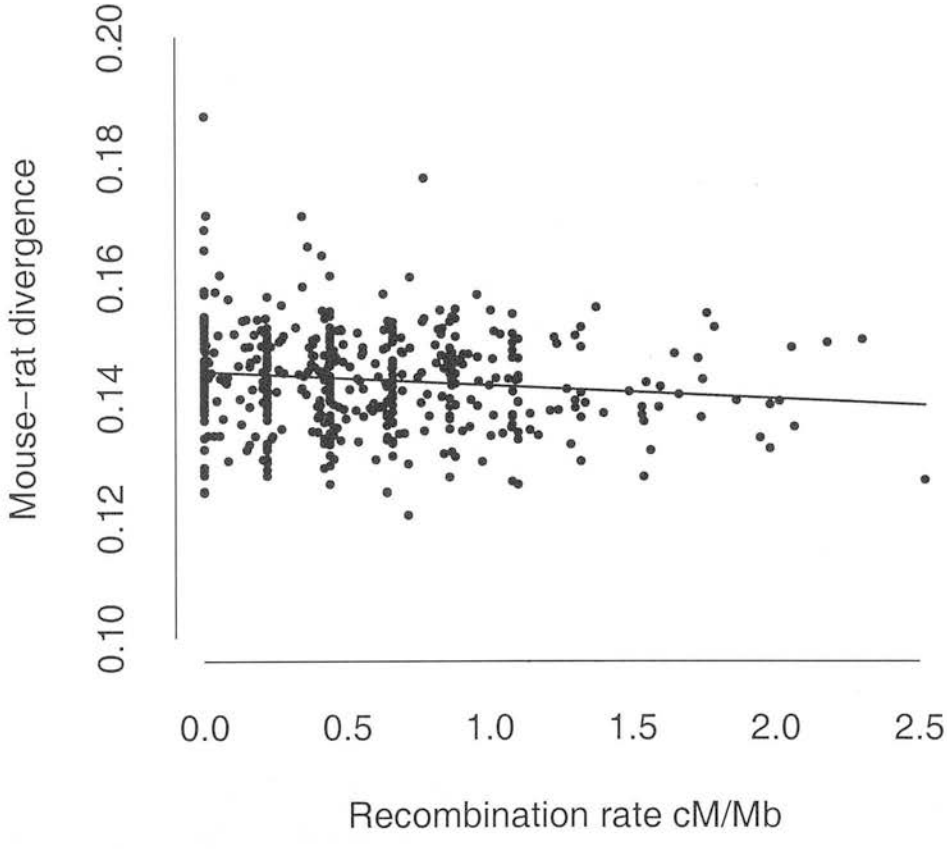

Figure 3.12.: The relationship between substitution rates at all sites and mouse recombination rate averaged across 5Mb windows. The equation of the regression line shown was estimated as $y = 0.144 - 0.002x$.

One implication of a subchromosomal mutational scale is that the process or processes that create point mutations could be expected to vary across similar scales. One candidate for such a driving process is recombination. Recombination

rates have been previously shown to covary with neutral substitution rates in ancestral repeats (Hardison et al., 2003). It is also known that recombination rates in humans are significantly correlated with GC content, probably as a result of biased gene conversion (BGC;Kong et al. 2002; Meunier and Duret 2004), although the contribution of BGC to human evolution has been debated by some (Comeron, 2006). Recent results from the highly recombining human pseudoautosomal region provide some evidence that recombination may have an effect on the neutral mutation rate (Perry and Ashworth, 1999; Filatov, 2004). In order to investigate the possibility that recombination rates are related to substitution rates, mouse recombination rate data was collected from a recent comparative study (Jensen-Seaman et al., 2004). These data consist of estimates of local recombination rate in 5Mb windows across the mouse genome. Average substitution rates were estimated for each of these windows from the data. However, there is little evidence for a strong relationship between mouse recombination rates and mouse-rat divergence; the slope of the regression line of substitution rates on recombination rates is approximately zero (Figure 3.12). If recombination drives mutation in murids then these data suggest that the relationship is not straightforward, at least on a genome wide level. This conclusion is supported by recent work suggesting that the relationship between recombination rate and nucleotide substitution is, at best, moderate (Huang et al., 2005). Furthermore, some studies have suggested that the majority of recombination events in humans occur in a comparatively small proportion of the genome (McVean et al., 2004; Crawford et al., 2004). Assuming recombination 'hotspots' also occur in murids, the lack of an observed relationship may be explained, in part, by this effect. For example, if recombination rates vary over scales of kilobases, as opposed to megabases, then any relationship between mutation and recombination may be obscured by averaging recombination rates over large genomic distances. In addition, rapidly changing recombination rates will cause problems in deciphering the true nature of any relationship between mutation and recombination, as the recombination is typically measured over much shorter timescales than the substitution rate.

One problem to which these data are potentially susceptible is that of gene conversion in repetitive sequence. It has been shown recently that some gene conversion occurs in young Alu repeats (Roy et al., 2000). If gene conversion is biased in the direction of the ancestral state, then this will produce a negative correlation between nucleotide divergence and the rate of conversion. The

distributions of repeat age within SINEs (results not shown) would suggest that the murid equivalents of Alus (B1 elements) differ from the other families of SINE in that there are a small proportion of B1 elements which are younger than other SINE elements. This could suggest that many B1 elements used were retrieved from low mutating regions or that biased gene conversion (BGC) towards the ancestral repeat is occurring. If the BGC is occurring on a substantial scale then there is little that can be done to remove this effect from the data, short of locating those elements which are ancestral in a more highly diverged species, for example human, to minimise the proportion of young B1s in the dataset. However if gene conversion is occurring in some B1s in this dataset, it appears to have a small effect on the results. The pattern of autocorrelation is practically unchanged if B1s are entirely removed from the dataset as is the ratio of intra- to inter chromosomal substitutional variation. In addition, previous analyses have concluded that gene conversion in repetitive DNA appears to have relatively small effects on neutral substitution rates at the genomic scale (Makova et al., 2004).

It has been shown that the scale of mutational similarity in murids extends from 100kb to 15Mb and that the 'unit' of mutational variation is no larger than 1Mb. At the 1 megabase scale, the results of this work suggest that there exists just under an order of magnitude more variation in mutation rates within chromosomes than among chromosomes. These results have important consequences for the study of the processes driving mutation and identification of functional noncoding DNA using comparative genomic methods.

4. Selective Constraint and Deleterious Mutation in Murid Noncoding DNA

The experimental design and implementation of work described in this Chapter were carried out by myself and Prof. Peter D. Keightley. All data collection was split evenly between myself and Prof. Keightley. All data analysis in this Chapter was performed by myself. This work has been published (Keightley and Gaffney, 2003).

4.1. Introduction

Mammalian genomes are characterised by substantial quantities of noncoding DNA (IHGSC, 2001; IRGSC, 2004; IMGSC, 2002). For example, the rodent genome is roughly 98.5% noncoding, approximately 40% of which is made up of the remnants of transposable elements insertions and other repetitive DNA types, such as satellite DNA (Figure 4.1). Whilst the majority of the repetitive portion of the murid genome is unlikely to be functional, the function of the remaining "unique" or "single-copy" noncoding DNA remains unknown. The aim of this chapter is to quantify selective constraint in some of the noncoding fraction of the rodent genome. The analysis was confined to noncoding DNA located within or nearby to protein-coding loci.

This study was facilitated by the availability of well-annotated genomic data from two closely related species, mouse and rat. These data enabled two important issues faced by previous analyses to be addressed. Firstly, because the majority of mammalian noncoding DNA is evolving at a significantly higher rate than protein-coding sequence, alignment of noncoding DNA between two evolutionarily distant species, such as human and mouse, is problematic. This is less of a problem between two closely related sister species, such as mouse and rat. Secondly, the availability of whole genome sequences means that the estimate of constraint at a locus can be calibrated with a reliable estimate of the local mutation rate. This is important because mutation rates are known to vary across the genome (Wolfe et al., 1989; Smith et al., 2002; Ellegren et al., 2003;

Gaffney and Keightley, 2005).

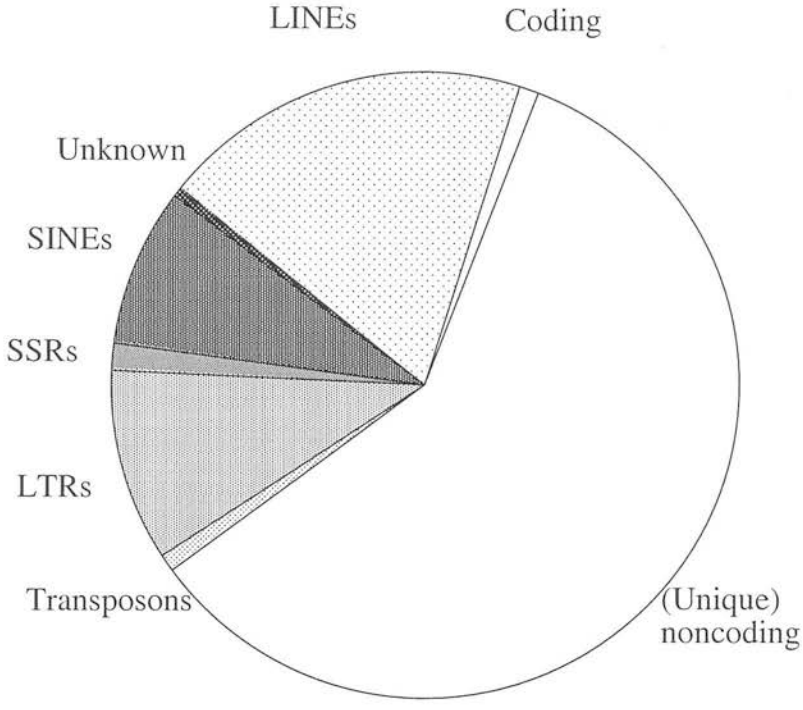

Figure 4.1.: Different types of noncoding DNA in the mouse genome. Abbreviations correspond to Long INterspersed Elements (LINEs), Short INterspersed Elements (SINEs), Long Terminal Repeat retrotransposons (LTRs) and Simple Sequence Repeats (SSRs). Data from the International Mouse Genome Sequencing Consortium (IMGSC, 2002).

The approach to calculating evolutionary constraint used an adaptation of a method first suggested by Kondrashov and Crow (1993) which has previously been applied to coding sequence in a variety of species (Eyre-Walker and Keightley, 1999; Keightley and Eyre-Walker, 2000) and to noncoding DNA in *Drosophila* (Halligan et al., 2004). The fundamental prerequisite of this method is a reliable estimator of the neutral mutation rate. Many previous studies have assumed that, in mammals at least, substitutions at the third position of four-fold degenerate codons are selectively neutral and are, therefore, a good estimator of the neutral mutation rate. However, weak selection at synonymous sites is known to occur in some prokaryotes (Hartl et al., 1994) and *Drosophila* (Akashi and Schaeffer, 1997; Akashi, 1997). Some recent studies have suggested that synonymous sites in mammals may also be under weak selection (Chamary and Hurst, 2005, 2004; Bustamante et al., 2002; Lu and Wu, 2005). Indirect evidence of non-neutral evolution at synonymous sites was also

found in this study (*see* Figure 4.3) and these sites were, therefore, rejected as accurate estimators of the local neutral mutation rate. Instead, it was assumed that those sites that appear to evolve the fastest are those that are least influenced by the action of selection. Using this criterion, it was determined that substitution rates at intronic sites, excluding intron 1 and outside regions known to be important in splicing, were the most accurate estimator of the neutral mutation rate. Estimating the local mutation rate from such sites allowed the calculation of the expected numbers of mutations within adjacent noncoding DNA. Comparison with the observed numbers of substitutions gives the number of mutations with effects deleterious enough to be rapidly removed by selection. This method assumes that beneficial mutations occur infrequently enough to be ignored. Providing this is a reasonable model of how noncoding DNA evolves, this method allowed estimation of (i) the location of selectively constrained noncoding DNA and (ii) and contributions mutations occurring in these regions to the genomic deleterious mutation rate, U.

4.2. Materials & Methods

Data Collection

Random regions of the mouse genome were picked and a mouse-rat gene orthologue selected within the sampled region. Selection of an orthologue was primarily based upon the strength of supporting experimental evidence, i.e., presence/absence of fully characterised, manually curated mRNA. Initially, conservation of intron number was used as one criterion of orthology. Thus, 200 of the gene orthologues sampled were chosen only if the number of introns was equal between mouse and rat and without regard to the proximity of the next nearest coding sequence upstream and downstream. For the remaining 100 gene orthologues, the constraint upon conserved intron number between mouse and rat was relaxed. In addition, the final 100 loci sampled were selected providing they were more than 6kb distant from the nearest coding sequence in either direction. This last criterion was implemented in order to increase the number of intergenic sites in the final dataset. Coding sequence was initially extracted and aligned using ClustalW (Thompson et al., 1994). If the resulting alignment was ambiguous, the locus was rejected. Otherwise, the total coding sequence, up to 6kb of 5′ and 3′ intergenic sequence and up to three introns (always including the first and last introns, and a randomly selected intermediate) were extracted from each genomic contig. In all cases, the NCBI mouse annotation was used

to determine the start/stop codon and intron/exon boundaries. In cases where a non-first intron exceeded 2kb in length, 1kb from each end of the intron was analysed. Up to 12kb from either end of intron 1 was sampled. All noncoding sequence was aligned using MCALIGN (Keightley and Johnson, 2004).

CpG Islands

CpG islands are areas of the genome, implicated in gene regulation, that are GC rich and exhibit an above average representation of CpG dinucleotides (Bird, 1986). In such regions, CpG sites tend to be unmethylated and potentially not subject to the accelerated mutation rate typical of CpG dinucleotides elsewhere in the genome. All CpG dinucleotides located within putative CpG islands were therefore treated as non CpG sites for estimation of the deleterious mutation rate. Putative CpG islands were determined using the CpG report tool in the European Bioinformatics Institute EMBOSS software suite (Rice et al., 2000). An island was reported if the observed to expected ratio of CpG dinucleotides exceeded 0.6, in a region of high (>0.5%) GC content, for ten successive 100bp windows (Gardiner-Garden and Frommer, 1987). Only islands longer than 200bp were recorded.

Masking Microsatellites and Potential Nonhomologous Regions

Any parts of the alignments containing repetitive microsatellite elements were excluded from the analysis, since evidence from *Drosophila* and humans indicates these evolve differently from nonrepetitive sequence (Calabrese and Durrett, 2003). Areas surrounding microsatellites often exhibited high non-homology and were also removed. In addition, all sequences were examined using a 40bp moving window and any region which, when viewed with the window, produced a "peak" of obviously non-homologous sequence was subsequently excluded.

Analysis

Neutral Standard

One of the prerequisites of the method of Kondrashov and Crow (1993) is a reliable estimate of the underlying neutral mutation rate. Although synonymous sites have frequently been used to estimate this rate (e.g. Eyre-Walker and Keightley, 1999; Keightley and Eyre-Walker, 2000), there is evidence to suggest that synonymous sites may be under weak selection in mammals (Chamary and Hurst, 2004, 2005; Parmley et al., 2006;

Kondrashov et al., submitted). In addition, previous work (Majewski and Ott, 2002) has suggested that intron 1 may contain a moderately high frequency of functional sequence and may, therefore, be subject to purifying selection. For the purposes of this analysis it was assumed that intronic DNA, excluding all first introns and splice sites, was evolving neutrally and a reliable estimator of the local neutral mutation rate.

CpG Effects

In mammals, hypermutable CpG dinucleotides account for a disproportionately large number of the substitutions observed between species (Arndt et al., 2003b). This poses a number of problems for the accurate estimation of selective constraint. Due to their elevated mutation rate, even small differences in the frequency of CpG dinucleotides can substantially alter the number of mutations occurring in a sequence. Thus, if the frequency of the CpG dinucleotide is higher (lower) in the "neutral" sequence than in the sequence of interest, constraint will be overestimated (underestimated) in the latter. This bias will be determined by the level of hypermutability of CpG sites and the magnitude of the difference in CpG frequency between the two compared sequences. Small differences in CpG frequency were evident between the "neutral" intronic (%CpG = 0.0095), first intron (%CpG = 0.012) and intergenic sequences (%CpG = 0.0133). Despite the apparently small magnitudes involved, simulations showed that these differences introduce considerable bias into the estimation of constraint (Figure 4.2) . In order to avoid introducing such a bias, the influence of CpG-derived mutation were removed, as much as possible, from the analysis, although necessarily CpG mutations were considered in the estimation of U. It was determined by simulation (*see* Chapter 2) that excluding 'CpG-prone' sites (i.e. sites preceded by C or followed by G) is robust method of excluding most CpG mutation. Simulations also showed that this improved the accuracy of estimation of constraint (Figure 4.2) in noncoding DNA. All estimates of constraint were, therefore, calculated at non CpG-prone sites.

Estimating Constraint

Constraint was estimated by extending the method developed by Kondrashov and Crow (1993). Firstly, each of the six separate pairwise substitution rates was calculated in the "neutral" intronic sequences. The product of each substitution rate and the number of appropriate sites (M_i for rate k_i; e.g. all A/A, A/T or T/T pairwise sites for the pairwise AT substitution

rate, k_{AT}) in the sequence of interest gives an estimate of the number of substitutions, E_i, expected under the null hypothesis of neutral evolution at a locus. Summing across all possible pairwise substitutions gives the total number of substitutions expected under neutrality

$$E = \sum_{i=0}^{6} k_i \hat{M}_i \tag{4.1}$$

Constraint, C, is calculated using the ratio of observed (O) to expected substitutions as follows:

$$C = 1 - \frac{O}{E} \tag{4.2}$$

Deleterious Mutation Rate

The contribution of mutations in noncoding regions to the overall deleterious diploid genomic mutation rate, U, was also calculated. Due to the large difference in CpG and non CpG mutation rates, the total deleterious mutation was estimated taking account of both rates separately. However, since it has been shown that estimating the substitution rate at CpG sites is seriously biased for small pairwise divergences (Chapter 2), instead the local CpG-prone (μ_{CG}) and non CpG-prone (μ_{nCG}) mutation rates were estimated from the appropriate fastest evolving intron sites for each locus. Each rate was estimated correcting for multiple hits using Kimura's two parameter model (Kimura, 1980). The deleterious mutation rate was estimated in segments over all loci. For each segment the number of deleterious mutations in segment i at locus j was determined as the product of the estimated local CpG-prone or non CpG-prone mutation rate, the number of CpG-prone (CGp_{ij}) or non CpG-prone ($nCGp_{ij}$) sites in a segment and the estimated mean constraint of that segment (C_i). This method assumes constraint at CpG-prone and non CpG-prone sites is equal. Mutation rates per nucleotide site were averaged and numbers of each different type of site were summed over all j sampled loci, over i successive segments of length l_i (200 bp for intergenic sequence and the 5′ region of intron 1; 6bp and 30bp for the 5′ and 3′ intronic splice control regions respectively). For each segment, the contribution to U was weighted by the number of sites l_i and the fraction of loci, P_i, that contain segment i as follows:

$$U = Z \sum_{i=1}^{segments} P_i l_i \frac{\sum_{j=1} \mu_{CG\,j} CGp_{ij} C_i + \mu_{nCG\,j} nCGp_{ij} C_i}{\sum_{j=1} CGp_{ij} + nCGp_{ij}}$$

where Z is a constant to scale from mutations per nucleotide site to mutations

per diploid genome per generation, as follows:

$$Z = \frac{2 \times 26512(loci^{\dagger}) * 0.5(\texttt{generations/year}^{\ddagger})}{2 \times 13 * 10^{6}(\texttt{divergence, time}^{\S})} \tag{4.3}$$

† (Hubbard et al., 2005), ‡ (Keightley and Eyre-Walker, 2000) § (Jaeger et al., 1986)

In the case of CpG islands, it was assumed that the CpG-prone mutation rate was equivalent to the non CpG-prone rate, (i.e. μ_{nonCGp}). Values for P_i were calculated from the first 200 loci, for intergenic regions which were sampled irrespective of proximity to other coding sequence, or from all loci, for introns.

Simulations

In order to assess the impact of CpG hypermutability upon the method of estimating selective constraint, a simple simulation was implemented. All "neutral" intronic, first intron and intergenic mouse DNA were concatenated into a single sequence. Using this as the ancestral sequence, a two-branch phylogeny was simulated and both lineages evolved under a CpG hypermutability mutation model. Each lineage was evolved until the point at which every site in each lineage had experienced 0.08 substitutions on average, approximately the same as the mouse-rat silent site divergence. The "neutral" intronic portion of the concatenated sequence was simulated to evolve neutrally and the intergenic and intronic portions were simulated to be under a selective constraint of 0.1. Selective constraint in the simulated first intron and intergenic sequences was estimated using substitution rates estimated at all and non CpG-prone sites. CpG hypermutability was varied from 1 to 20-fold greater than the non CpG mutation rate and estimates were averaged over 100 replicates.

4.3. Results

A total of 300 orthologous loci were extracted from the whole genome assemblies of mouse and rat in GenBank. This provided 20 kb of coding sequence, 2.3 Mb of intergenic DNA (1.2 Mb from 5′ region, 1.1 Mb from 3′ region) and 1.5 Mb of intronic DNA (1.0 Mb first intron, 0.5 Mb non-first intron).

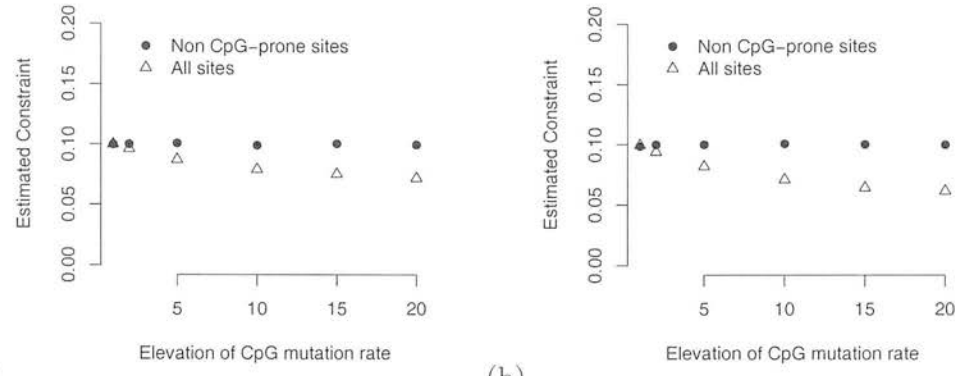

Figure 4.2.: Constraint estimated in sequences evolved under a CpG hypermutation model. Constraint of 0.1 was simulated in real mouse first intron (a) and intergenic (b) sequences and calculated relative to neutrally evolved non-first introns. Constraint was estimated using rates calculated at non CpG-prone (•) and all sites (△).

Simulations

The results of the simulations are presented in Figure 4.2. It is clear that, even the marginal differences in CpG frequency between fastest evolving intron sites and other noncoding regions have a serious impact on the estimation of constraint. CpG frequency is lowest at fastest evolving intron sites (0.0095), compared with intron 1 (0.0120) and intergenic DNA (0.0133). When substitution rates at all sites are considered, constraint is underestimated in simulated first introns and intergenic DNA because the slight differences in CpG frequency increase the observed substitution rate. Assuming 10-fold CpG hypermutability, constraint is underestimated by $\sim 20\%$ in simulated first introns and by $\sim 30\%$ in simulated intergenic DNA, simply due to differences in CpG frequency. This results also show that excluding CpG-prone sites appears to be a reliable method of removing the effects of CpG hypermutation and allows more accurate estimation of selective constraint.

Substitution Rates at Fourfold and Intronic Sites

The results show that fourfold degenerate non CpG-prone sites are evolving significantly slower than non splice site intronic non CpG-prone sites (Figure 4.3). Five of the six pairwise substitution rates are significantly higher in introns than at fourfold sites ($P<0.003$ or less). Only substitutions between G and C are indistinguishable between introns and synonymous sites ($P=0.58$; Figure 4.4 b). Significant transition/transversion bias is also apparent at both intronic and

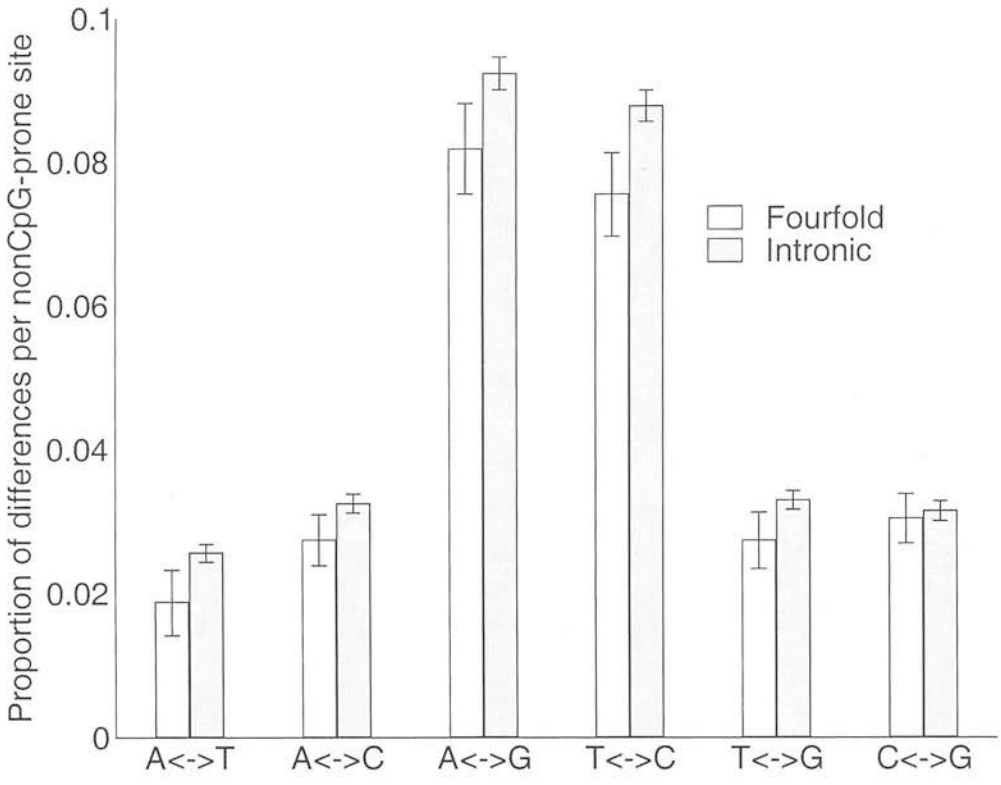

Figure 4.3.: Proportions of differences at non CpG-prone fourfold degenerate and intronic sites. Error bars show 95% C.I.

fourfold sites.

Constraint

Figure 4.5 shows the selective constraint of intergenic DNA plotted against distance from the 5′ and 3′ ends of the coding sequence. Constraint becomes statistically indistinguishable from zero within approximately 4kb of both the start and stop codons. The level of constraint in the immediate vicinity of the coding sequence is moderately high, with approximately one in every three new mutations within 200 bp of the start and stop codons being removed by purifying selection.

Constraint of the donor (5′) and acceptor (3′) sites of non-first introns and the region immediately proximate to them is high (Figure 4.6), and approaches that observed at nondegenerate coding sites. This pattern reflects the importance of these sites for accurate intron excision from the pre-mRNA molecule (Li, 1997).

The 5′ region of intron 1 (Figure 4.7) is also moderately selectively constrained to a distance of approximately 1700bp from the 5′ splice site. One explanation for this pattern is that motifs important for gene transcription control are

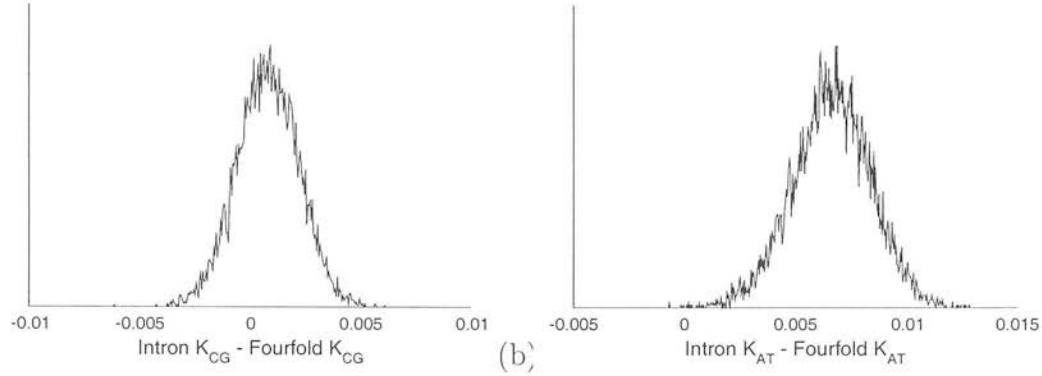

Figure 4.4.: Frequency distribution of the difference in mean C↔G (a) (K_{CG}) and A↔T (b) (K_{AT}) substitution rates at intronic and fourfold sites, estimated by bootstrapping the data by gene 10,000 times.

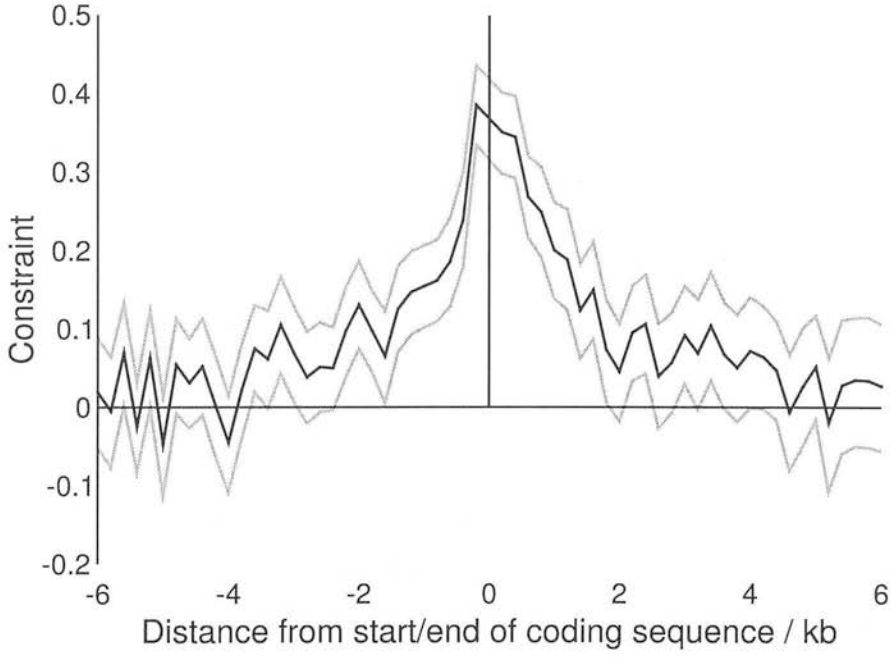

Figure 4.5.: Average constraint of 200 bp blocks up to 6kb upstream and downstream of the coding sequence. 95% C.I. of the estimate of constraint in each block is shown in grey.

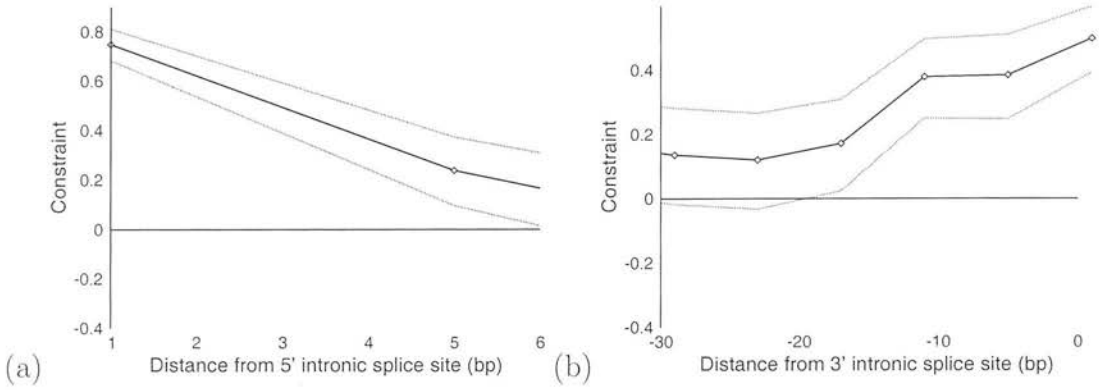

Figure 4.6.: Average constraint of 4bp blocks of the 5′ (a) and 3′ (b) splice control region of non-first introns. 95% C.I. of the estimate of constraint in each block is shown in grey.

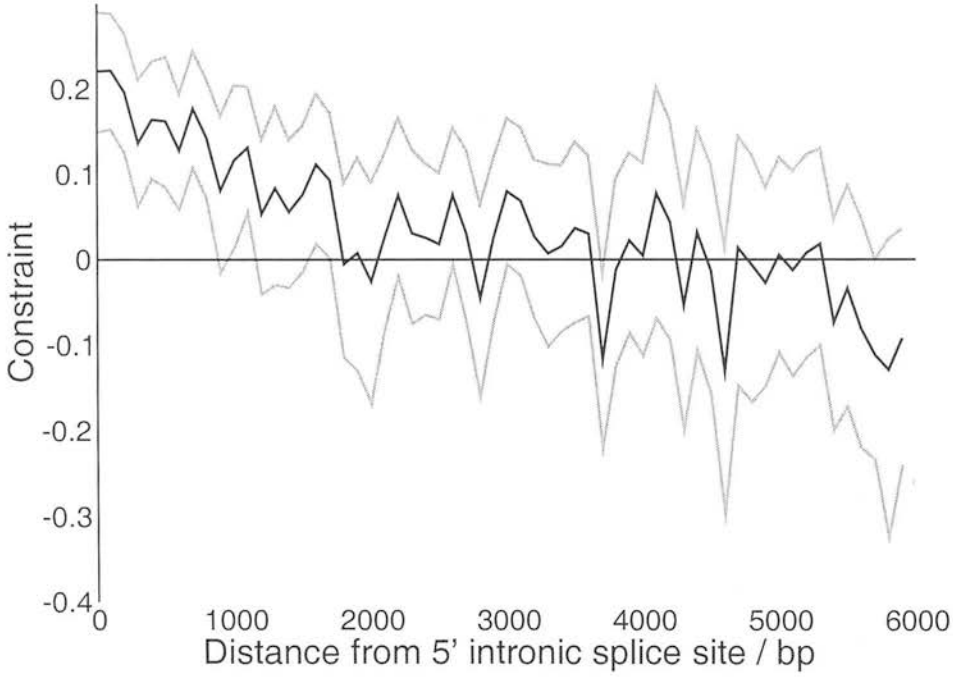

Figure 4.7.: Average constraint of 100bp blocks downstream of the 5′ splice site of intron 1. 95% C.I. is shown in grey.

preferentially located within intron 1. In contrast, no significant constraint was found outside of the splice regions in the 3′ region of intron 1 (Figure 4.8) which appears to be similar to other non-first introns.

Deleterious Mutation Rate

Estimates of the genomic deleterious mutation rates in coding and noncoding DNA are presented in Table 4.1. Although selective constraint of noncoding regions is uniformly lower than that in coding sequence, the relatively large

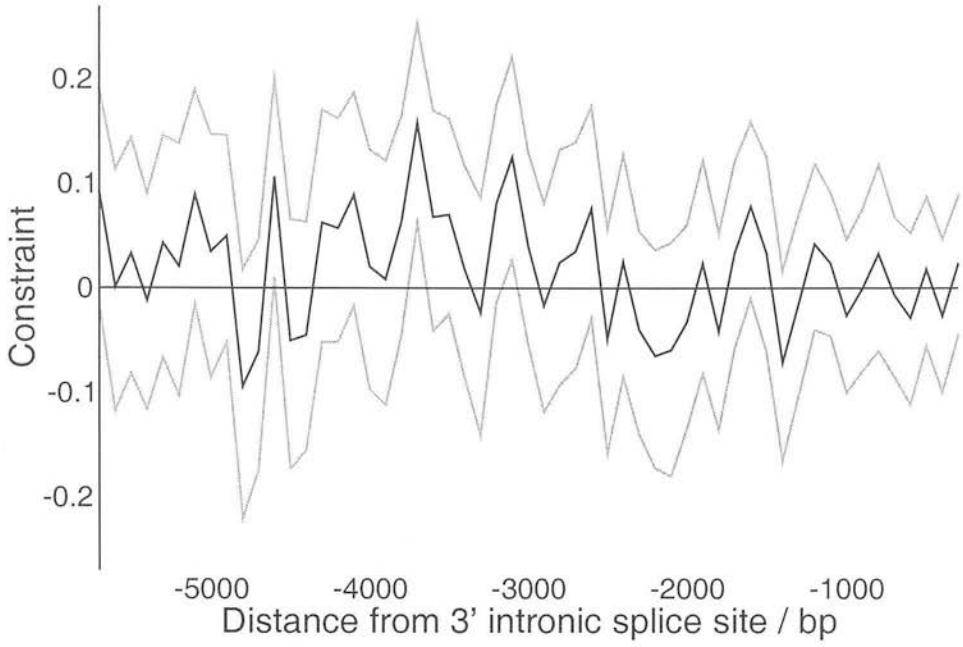

Figure 4.8.: Average constraint of 100bp blocks upstream of the 3′ splice site of intron 1. 95% C.I. is shown in grey.

quantities of noncoding DNA in the murid genome mean that just under half of all deleterious mutations occur within noncoding DNA, most of which occur in 5′ and 3′ intergenic regions.

4.4. Discussion

The results of this study suggest that, in the murid genome, much of the noncoding DNA adjacent to protein-coding loci is under moderately high selective constraint. In intergenic DNA the magnitude of this constraint decays with distance from the coding sequence, reaching zero at approximately 6kb upstream/downstream from the start/stop codons. The results also show substantial constraint in the 5′ region of intron 1. The large quantity of noncoding DNA present in the murids means that the estimated genomic deleterious mutation rate, U, in noncoding DNA is approximately equal to that in protein-coding regions.

Excluding CpG mutations, fourfold degenerate synonymous sites are evolving more slowly than fastest evolving intronic sites. This finding is supported by recent studies which suggest a reduced substitution rate at fourfold sites in humans and rodents (Bustamante et al., 2002). However, although these results may provide indirect evidence for weak selection at murid synonymous sites, it would be unwise to draw many conclusions from this result. It is

Table 4.1.: Average constraint and genomic deleterious mutation rate in each different site class. Constraint was calculated as in Eq. 4.2. It was assumed 75% of 1,500 sites per coding sequence were nondegenerate. Estimates of constraint at intron splice sites were estimated from 6bp and 30bp for 5′ and 3′ regions respectively, assuming an average of 7.4 introns per locus. Constraint in the 5′ region of intron 1 excluded estimates from intron 1 splice sites.

Region	Sites / locus	Constraint (SEM)	U genome^{-1} generation^{-1}
Coding	1125	0.87 (0.009)	0.1610 †
5′ splice region	44.4	0.721 (0.029)	0.0049
3′ splice region	222	0.289 (0.0247)	0.0090
Intron 1, 5′ end	3336	0.067 (0.0179)	0.0073
5′ Intergenic	5596	0.094 (0.0135)	0.0542
3′ Intergenic	5271	0.120 (0.0155)	0.0575
Total			0.2939

†As in (Keightley and Eyre-Walker, 2000), but assuming 25,612 loci per genome (Hubbard et al., 2005)

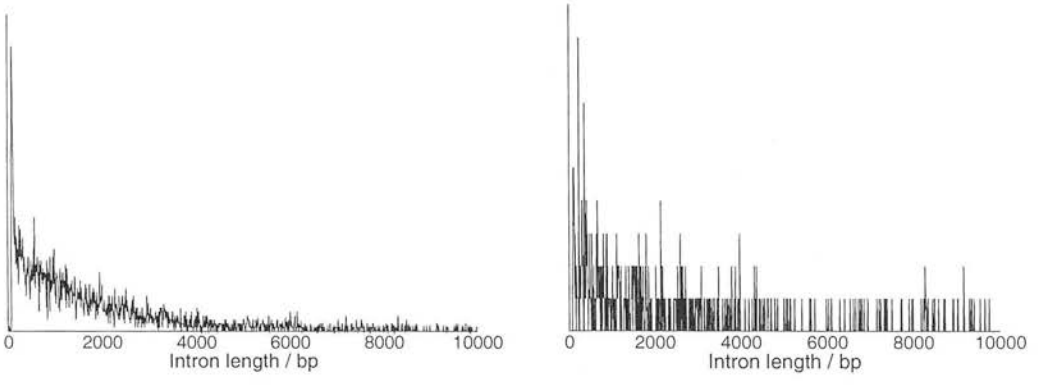

Figure 4.9.: Frequency distribution of intron lengths for non-first introns (a) and first introns (b).

possible that this result reflects some inherent mutation bias at fourfold sites which has not been accounted for. This result could also reflect the action of biased gene conversion (Marais, 2003), although it is unlikely that this would affect fourfold and intronic sites differently. Nonetheless, evidence for selection at mammalian synonymous sites is accumulating (Chamary and Hurst, 2005, 2004; Parmley et al., 2006) and recent work has suggested that selection may increase the GC content at synonymous sites (Kondrashov et al., submitted). This is one scenario for which there is also some indirect evidence of in this

study, as fourfold codons in the current dataset are substantially biased towards those codons ending in G or C (~74%). It is perhaps difficult to imagine how such substantial compositional skew could persist by mutational bias alone, particularly considering the comparatively low GC content of most of the rest of the genome (~40%; IMGSC 2002), although this could also reflect the activity of transposable elements.

The symmetrical decay of constraint in both 5′ and 3′ regions indicates approximately equal intensity of purifying selection at both the ends of the coding sequence. In addition, the relatively slow rate of decay of constraint with distance from the coding sequence is perhaps unexpected, indicating that other elements important to normal gene function may lie some distance from the coding sequence. The proportion of constraint in these regions that is specifically due to conservation within the 5′ and 3′ Untranslated Regions (UTRs) has not been estimated in this study. If recent results from *Drosophila* are any indication (Andolfatto, 2005) then it is likely constraint in UTRs is substantial.

This analysis has also revealed that constraint in the first intron is higher than within introns. This would suggest that functional elements, most likely involved in gene transcription control, are preferentially located within the first intron. This result is supported by other analyses which have shown that the frequency of repetitive Short Interspersed Nuclear Elements (SINEs) in human first introns is much lower than in the rest of the genome (Majewski and Ott, 2002) . If intron one is the location of motifs which are functionally constrained, such as transcription-factor binding sites, transposable element insertions are more likely to be deleterious in this region and removed by selection. The evidence would seem to suggest, therefore, that it is beneficial to have elements involved in gene regulation located near to the transcription start point, although why this is the case remains unclear.

The method used here to estimate selective constraint relies on two assumptions. Firstly, it was assumed that neutrally evolving sequence with which to calculate the expected rate of evolution in the absence of purifying selection was correctly identified. There are at least two reasons why this assumption may be invalid. If substantial adaptive evolution was occurring in non-first introns, then the rate of substitution would exceed that expected under neutrality. In this scenario selective constraint would be overestimated in the

regions studied. The frequency of adaptive evolution in mammalian noncoding DNA is not clear at the present time. It has recently been suggested that a substantial fraction of new mutations in *Drosophila* noncoding DNA may have been driven to fixation by positive selection (Andolfatto, 2005). However, it is becoming clear that the structure and evolutionary history of the *Drosophila* genome may be markedly different from that in mammals. A study of the intergenic DNA upstream and downstream of human and chimp orthologous genes has shown that the fraction of adaptive substitutions appears to be very low (Keightley et al., 2005b). Given that the human-chimp ancestral genome is likely to be more similar to the murid genome than that of *Drosophila* it seems unlikely that a large fraction of murid intronic DNA is undergoing adaptive evolution. Perhaps a more likely problem with the assumption of neutrality is that there is some level of selective constraint in those introns that were assumed to be evolving neutrally. In this scenario constraint will be underestimated. A potential resolution of this is to calibrate estimates of constraint with sequence with which a better biological case for neutrality can be made, such as repetitive DNA. This question has been addressed in detail elsewhere (Chapter 5).

The second assumption of the method employed here is that the mutation rate in the "neutral standard" is the same as that in the sequences of interest. Although the fastest evolving introns were, by definition, never very distant from the intergenic or first intronic DNA, this is no guarantee that mutation rates in both were identical. By excluding CpG-derived mutations one factor likely to induce substantial short-range fluctuation in mutation rates has been removed. However, more detailed information is required about the scale(s) over which mutation rates vary before the validity of the assumption of equal mutation rates can be assessed. Again, this question has been addressed elsewhere (Chapter 3).

The estimate of U indicates that, on average, murids experience a single deleterious mutation event per genome every 2.5 generations. This is too low to explain the evolution of sexual reproduction under the mutational deterministic hypothesis (Kondrashov, 1988). However, U has almost certainly been underestimated for a number of reasons. As mentioned, if the "neutral standard" introns are selectively constrained, the estimate of U will be too low. In addition, whilst the sample size used was relatively large, it was necessarily constrained by the computational intensity of the alignment protocol. There may be other significantly constrained noncoding regions that have not been

sampled, both within intergenic regions and larger introns. Thus, U may have been underestimated by excluding weakly constrained areas outside the sample of noncoding DNA. Nonetheless, the distribution of intron lengths in the data (Figure 4.9) indicates that at least half the total length of over 80% of those non-first introns sampled in this analysis has been included. Likewise approximately 87% of sites within first introns have been sampled.

U may also have been underestimated in noncoding regions by excluding more distant intergenic regions. Although it has been generally assumed that the majority of important regulatory features lie within 2kb of the coding sequence (IMGSC, 2002), recent work has revealed the extent of highly conserved nongenic (CNGs) sequences whose frequencies across the genome are negatively correlated with that of protein-coding loci (Dermitzakis et al., 2004; Bejerano et al., 2004; Gaffney and Keightley, 2004). Quite a high proportion of CNGs have potentially been excluded due to sampling of relatively nearby intergenic sequence. The selective constraint in a much larger sample of murid intergenic DNA has been estimated elsewhere in this thesis (*see* Chapter 5). Although the genomic deleterious mutation rate in murid CNGs has been estimated to be moderately high (0.15 Keightley et al. 2005a), however, this does not increase the estimate of U above one. A final reason that the estimate of U may be too low is that the effects of indel mutation have been neglected in this analysis, which are expected to further contribute to the deleterious mutation load.

A problem with the estimation of U is that much weight in the calculation is lent to constants such as genes per genome, generations per year and species divergence times which frequently rely on controversial assumptions. The *Mus-Rattus* divergence time is hotly debated, and whilst palaeontological estimates have indicated a date of approximately 13-14 Mya (Jacobs and Pilbeam, 1980; Jaeger et al., 1986), molecular estimates have ranged from 5-40 Mya (Wilson et al., 1977; Kumar and Hedges, 1998). Estimates of species divergence times from fossil data were used in this study to avoid potentially invalid assumptions of rate constancy in molecular studies. If the divergence times used are even slightly inaccurate, however, this will have a comparatively large effect on the estimate of U. Similar criticisms can be applied to estimates of gene number and to generation time.

As with any study of this kind, a certain sampling bias will also have an

influence on estimates of selective constraint, namely that sequences under comparatively weak evolutionary restrictions will, by definition, have diverged to a greater extent. Such sequences were less likely to be sampled in this study, given the criteria imposed to ensure alignment was as unambiguous as possible. As such, it is possible that the sample was biased towards more conserved loci. However, such a bias is unavoidable in any study which relies even partially upon the alignment quality as a criterion of orthology.

It is likely that a sizeable proportion of the constrained sites found in this study found are involved in gene regulation and expression control. This may indicate that a substantial proportion of new deleterious mutants are regulatory in nature. Given the likely downward bias of estimates of constraint these results do not allow a role for deleterious mutations in the maintenance of sexual reproduction for this species to be ruled out. In addition, larger scale analyses in *Drosophila* has indicated that the contribution of noncoding DNA to U can be even more substantial than observed here (approximately 3-fold greater than that in coding regions; Andolfatto 2005).

5. Genomic Selective Constraint in Murids

5.1. Introduction

As has been mentioned previously, protein-coding regions make up a rather small part of many mammalian genomes (IHGSC, 2001; IMGSC, 2002; IRGSC, 2004; ICGSC, 2005). The function of much of the remaining portion of the genome has remained, until relatively recently, almost entirely unknown.

One approach to this question involves the computational or experimental comparison of many segments of noncoding DNA from a target species with the entire genome(s) of a single, or multiple, outgroup species (Frazer et al., 2001; Dermitzakis et al., 2002, 2003; Bejerano et al., 2004). Such large scale comparative analyses have revealed a substantial amount of sequence that has remained highly conserved between multiple species, often across large periods of evolutionary time. Subsequent work has indicated that, whilst a number of these sequences may be undiscovered protein-coding genes or overlap with existing genes (Bejerano et al., 2004), the evidence does not support a protein-coding function for conserved regions in many cases (Dermitzakis et al., 2003, 2005). These regions have been referred to as conserved non-genic sequences (CNGs). In addition to their lack of similarity to known protein-coding sequences, CNG frequency across the genome appears to be negatively correlated with gene density (Gaffney and Keightley, 2004; Dermitzakis et al., 2004).

One problem with this approach is that, whilst large scale comparisons provide an estimate of the amount of conserved sequence between multiple species, an explicit null hypothesis regarding the functionality of such sequences, and the evolutionary pattern expected under this null is not clear. It is known that the substitution rate varies across the mammalian genome (Wolfe et al., 1989; Ellegren et al., 2003; Gaffney and Keightley, 2005) and thus the expected frequency with which sites are conserved over evolutionary time will be higher in some regions than others. In addition, there is considerable ascertainment bias involved in picking the most evolutionarily conserved sequences in a genome, although this can be quantified and corrected for by the use of independent

phylogenetic comparisons (Keightley et al., 2005a). An alternative approach to the question of the importance/functionality of noncoding DNA involves calibrating observed evolutionary rates in potentially functional regions using nearby sequences that are assumed to be evolving neutrally. This enables the estimation of selective constraint (i.e. the proportion of new mutations arising which are deleterious and are removed by selection) within a sequence of interest. Selective constraint in turn provides information about the genomic deleterious mutation rate, U, a parameter of crucial importance in population genetics theory.

In a previous study in murids, we adopted the latter approach and showed that regions immediately adjacent to genes appear to be more highly selectively constrained than those regions that lie further (up to approximately 6kb) into intergenic regions (Keightley and Gaffney 2003; Chapter 4). However there are a number of limitations to this previous analysis which make a direct comparison of these two findings difficult. Firstly, this previous study was, due to the computational limitations of alignment, confined to those intergenic regions located comparatively close to protein-coding loci. Thus, the evolutionary significance of noncoding DNA located large distances from genes was not assessed. In addition, this study used the fastest evolving sites in introns as an assumed "neutral standard" with which selective constraint in noncoding DNA could be estimated. The assumption of neutrality is crucial to the accurate estimation of selective constraint. If some intronic sequence is functional, then selective constraint in noncoding DNA, and the total genomic deleterious mutation rate, will have been underestimated.

The purpose of the current chapter is, therefore, to address some of the limitations of the previous study. A particular objective was to determine the relative importance of "adjacent" and "distant" intergenic noncoding DNA and to determine the validity of the assumption of intronic neutrality. To this end genomic regions with which a reasonable biological case for neutrality can be made were identified. In mammals at least, it is likely that transposable elements are the most reliable candidates for neutrally evolving sequence in the genome. The observed nucleotide substitution rates in repetitive DNA were used as an assumed neutral standard throughout. This allowed the relative evolutionary importance of different genomic noncoding regions to be addressed in detail.

5.2. Materials and Methods

Data collection

A list of known mouse peptide sequences was obtained from the ENSEMBL sequence database. This list consists of peptides that are, by definition, those which can be mapped to mouse-specific peptides in either the Swiss-Prot, RefSeq or SPTrEMBL databases (Hubbard et al., 2005). This means that all of the exons in the source data were supported by comparison with existing mouse proteins, cDNAs and ESTs. Those peptides that did not match an existing sequence in the NCBI RefSeq database were removed. Those peptides which were annotated in ENSEMBL as having multiple transcripts were also removed, due to the uncertainty of annotation of introns in alternatively spliced genes. Finally those peptides which were not listed as having a unique best reciprocal BLAST hit (UBRH) in the rat genome on the ENSEMBL website were removed. The remaining peptides were then mapped onto NCBI build 33.1 of the mouse genome using their RefSeq ids, and their coding sequences were extracted.

Putative rat orthologues of the mouse coding sequences were located by comparing the first and last exons of each mouse coding sequence to NCBI build 3.1 of the rat genome using BLASTN (Altschul et al., 1997). Mouse coding sequences were only compared with regions of the rat genome which are known to be syntenic, where synteny was as defined in Figure 4 of the IRGSC (2004). The size of the "flanking" sequences was defined to be half the distance upstream and downstream of the coding sequence to the next annotated mouse coding sequence. BLAST matches in the rat genome were accepted or rejected in the basis of a number of criteria. Firstly, unless both first and last exons had a single unique match on the same rat contig, the mouse "query" sequence was rejected. Multiple matches of both exons on multiple contigs were also rejected. Secondly, both first and last exon matches were rejected unless BLAST hits were matched on the same strand of the rat genome. Finally, matches of the first and last exons which were further than 1Mb apart on the same rat contig were also rejected. For those matched pairs of first and last exons which fulfilled these criteria, everything between the start of the upstream flank to the end of the downstream flank in both mouse and rat was extracted, using the matched exons in the rat genome as a reference. These sequences were then aligned using AVID (Bray et al., 2003). Ancestral repeats (i.e. those that were inserted into the last common ancestor of mouse and rat) were located in the alignment

using RepeatMasker (http://www.repeatmasker.org/). The aligned coding and intronic sequence was extracted, using the annotated mouse exons as a reference. Any mouse or rat sequence which did not have a valid start and stop codon, or included premature stop codons was excluded. The remaining coding sequences were realigned using CLUSTALW (Thompson et al., 1994) to align the amino acid sequences which in turn produced the DNA sequence alignments.

Alignment Masking

In order to minimise the possibility of nonhomologous sites contributing to estimates of divergence, a simple masking protocol was implemented. Two primary masking targets were identified. (i) Sections of alignments which were so divergent as to be unlikely to be homologous were located through the use of a sliding window of 40bp in size. Any region in which each of 30 or more consecutive windows showed a mean divergence greater than the threshold divergence of 30% was masked. The divergence threshold was set to be three standard deviations from the mean divergence of ancestral repeats (mean = 0.1596; s.d. = 0.0504). (ii) Regions which contained short aligned blocks surrounded by large gaps were also considered unlikely to be truly nonhomologous and were masked off. These regions were identified as one or more blocks of < 20bp in size, flanked by large gaps (> 40bp) in size. Any alignments which contained $\geq 75\%$ putatively nonorthologous sites as identified by this protocol were excluded in their entirety from further analyses. In addition to masking putatively nonorthologous sites, repetitive sequence (simple sequence, retroelements and DNA elements) present in the intron alignments were also masked using RepeatMasker. The reasons for this are twofold. Firstly, tandem repeats are known not to mutate by the same mechanism as single point mutations and could erroneously inflate or deflate estimates of divergence. Secondly, this study specifically addresses constraint within unique, nonrepetitive sequence.

Data Analysis

Nucleotide substitution rates were corrected for multiple hits according to the Tamura and Nei (1993) model. In order to investigate the influence of CpG-derived mutation on estimates of genome-wide substitution rates, nucleotide substitution rates were initially calculated at all sites and those sites that were not preceded by a "C" and/or followed by a "G" (non CpG-prone sites). In addition, it has been suggested that the level of methylation may differ

substantially between repetitive and nonrepetitive DNA (Meunier et al., 2005). If this is the case, CpG sites effectively mutate at different rates depending upon their location in the genome, and it is desirable to remove this effect as much as possible from the estimation of constraint. Substitution rates at linked sites are also autocorrelated across distances of $\sim$ 1 Mb (Gaffney and Keightley 2005; Chapter 3). All gene orthologues were therefore grouped into 1Mb blocks, according to their annotation on the mouse genome, to minimise the effects of autocorrelation of substitution rates on the estimation of standard errors and confidence intervals. Blocks were treated as single independent observations in the dataset. Substitution rates in different sequence types were estimated by summing across all annotated regions of interest (e.g. all non-first introns or intergenic transposable elements) within a block. Synonymous substitution rates were estimated at fourfold degenerate sites only. All standard errors and confidence intervals were calculated by bootstrapping the data over 1Mb block 1000 times.

In order to estimate selective constraint, a variation of the method of Kondrashov and Crow (1993) was employed, as in previous studies (Eyre-Walker and Keightley, 1999; Keightley and Eyre-Walker, 2000; Keightley and Gaffney, 2003). Each 1Mb block was treated as a single, independent observation. For each sequence class, observed substitution rates were compared to that expected under neutrality, where the neutral expectation was calculated using the substitution rate estimated, summing across all ancestral repeats within a block. One problem with this method is that different substitutions occur at different rates. When base composition varies between assumed neutrally evolving sequence and the sequence of interest, differences in the frequencies of each nucleotide can introduce error into the estimation of the expected evolutionary rate under neutrality. This was accounted for by estimating the expected substitution rates for different nucleotides separately. The model of Tamura and Nei (1993) is is described by three substitution rate parameters: the A$\leftrightarrow$G substitution rate (K_{AG}), the T$\leftrightarrow$C substitution rate (K_{TC}) and the transversion substitution rate (K_{TV}). The rate of substitution expected under neutrality was calculated as the product of each of the mean ancestral repeat substitution rates and the number of appropriate bases for that substitution type in the target sequence of interest. We have:

$$C = 1 - \frac{K_{AG\,i}N_{AG\,i} + K_{TC\,i}N_{TC\,i} + K_{TV\,i}N_{TV\,i}}{K_{AG\,ar}N_{AG\,i} + K_{TC\,ar}N_{TC\,i} + K_{TV\,ar}N_{TV\,i}}$$

where $K_{.i}$ denotes the substitution rate in the sequence of interest, $N_{.i}$ the number of sites of a certain type in the sequence of interest and $K_{.ar}$ the mean substitution rate across all ancestral repeats located in a block. In all cases, constraint was estimated at non CpG-prone sites, both to remove the considerable effects of differential CpG frequency between different sequence classes (e.g *see* Chapter 4), as well as to avoid the potential effects of differential methylation of repetitive and nonrepetitive DNA (Meunier et al., 2005). When investigating variation with sequence length and selective constraint, sequence length was always defined as the number of bases in mouse.

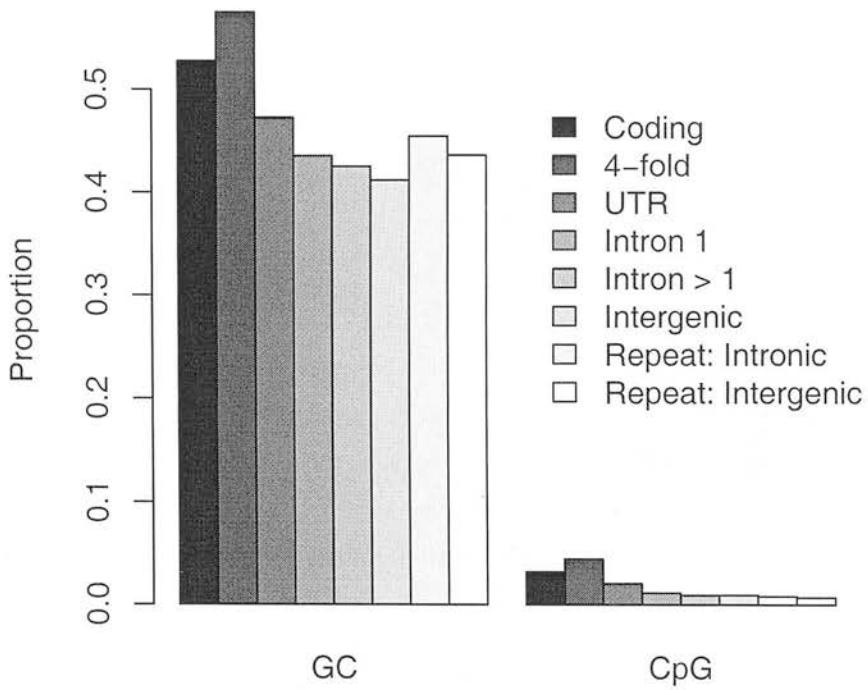

Figure 5.1.: Proportion of G/C bases and CpG dinucleotides in different mouse genomic sequence classes.

Simulations

The rate at which any one nucleotide mutates into another is known to vary (Li et al., 1984). Two notable examples of this would include transition/transversion bias and context-dependent mutation (e.g within CpG

dinucleotides). It is important to account for these processes when comparing evolutionary rates across sequences of varying base composition. In particular, it was necessary to determine to what extent varying base composition and a realistic mutation model could explain the patterns of substitution observed under the null hypothesis of no selective constraints. To do this, a simple simulation protocol was implemented. Sequences were divided into three approximate groups on the basis of their observed GC and CpG contents: fourfold degenerate, intronic/intergenic and ancestral repeat (Figure 5.1). In order to accurately reflect the variation in base composition between groups, simulated phylogenies were generated using real mouse sequence data. All mouse coding sequences, and a random sample of mouse intronic and repetitive sequences (of approximately the same length as that of the coding DNA used(8.5Mb)), were concatenated into a single sequence. This sequence was evolved along two, independent branches to produce a tree in which the probability of a nucleotide substitution at any site in either lineage was 0.08, on average. 80% of amino-acid-changing mutations in coding sequence were rejected. The remainder of sites (fourfold degenerate, intronic and repetitive) were allowed to evolve neutrally. A different random sample of intronic and repetitive sequence was collected for each replicate phylogeny. The mutation model implemented was derived from human polymorphism data collected by Prof. Peter Keightley from the Environmental Genome Project (EGP). 529 human-chimp gene orthologues were identified using a best reciprocal BLAST hits approach. All human introns for these genes available in the EGP database in the were extracted and aligned with AVID (Bray et al., 2003) to the appropriate chimpanzee outgroup, in order to assign polarity to each observed polymorphism. Chimpanzee sequence was derived from the draft chimpanzee genome sequence. The use of polymorphism data with a closely related outgroup enabled a relatively unambiguous assignment of the ancestral state at each aligned site. Using these data, the relative mutabilities at three site types (CpG, non CpG "G" or "C" and non CpG "A" or "T") were estimated. The relative probabilities, given that a mutation occurs at one of the site types, of a change to a GC or AT site were also derived. The mutabilities and relative probabilities used in the simulation are presented in Table 5.1. The transition/transversion ratio was set to one.

Assuming that the mutation model implemented accurately represents true murid mutation rates, simulated phylogenies should approximate the rates and patterns of nucleotide substitution expected under neutral evolution of fourfold

Table 5.1.: Mutabilities and relative probabilities of mutations used in the simulation for the three site types: CpG, non CpG GC and non CpG AT. Mutability is the probability of a mutation occurring at one the three site types. "→AT" is the relative probability of a mutation to either "A" or "T", similarly for "→GC", given that a mutation occurs.

	CpG	non CpG AT	non CpG GC
Mutability	0.816	0.078	0.106
→ AT	0.954	0.141	0.859
→ GC	0.046	0.794	0.206

sites, introns and transposable elements.

5.3. Results

The initial list of "known" ENSEMBL mouse genes contained 24560 peptides. Processing to remove those peptides which did not meet the selection criteria described in Section 5.2 left 8932 mouse genes. Upon comparison with the rat genome using BLASTN those matches which appeared to be invalid or matched a rat sequence which was not also a coding sequence (i.e. had premature stop codons) were excluded. This left a total of 6381 putative mouse-rat orthologous loci, which provided a total of 1.3 million fourfold degenerate sites. Excluding masked bases, a total of 85.9 Mb aligned intronic sequence and 139.4 Mb aligned intergenic sequence (of which 73.4 Mb was 5′ and 66.0 Mb 3′ intergenic) was extracted. The alignments also provided a total of 62.5 Mb ancestral repeat sequence, of which 20.1 Mb was located within introns and the remaining 42.4 Mb located within intergenic regions. The total aligned sequence was 293.7Mb, approximately 25% of the total alignable sequence between mouse and rat (Figure 7, IRGSC 2004).

Substitution rates in coding and noncoding DNA

The mean nucleotide substitution rate was estimated for each of a total of seven sequence classes: fourfold degenerate sites, Untranslated Regions (UTRs), first introns, non-first introns, intergenic DNA and intronic and intergenic ancestral repeats. The results of this analysis are presented in Figure 5.2 and show nucleotide substitution rates estimated at all sites and at non CpG-prone sites. It is clear that the latter rate is substantially less than the former for all sequence

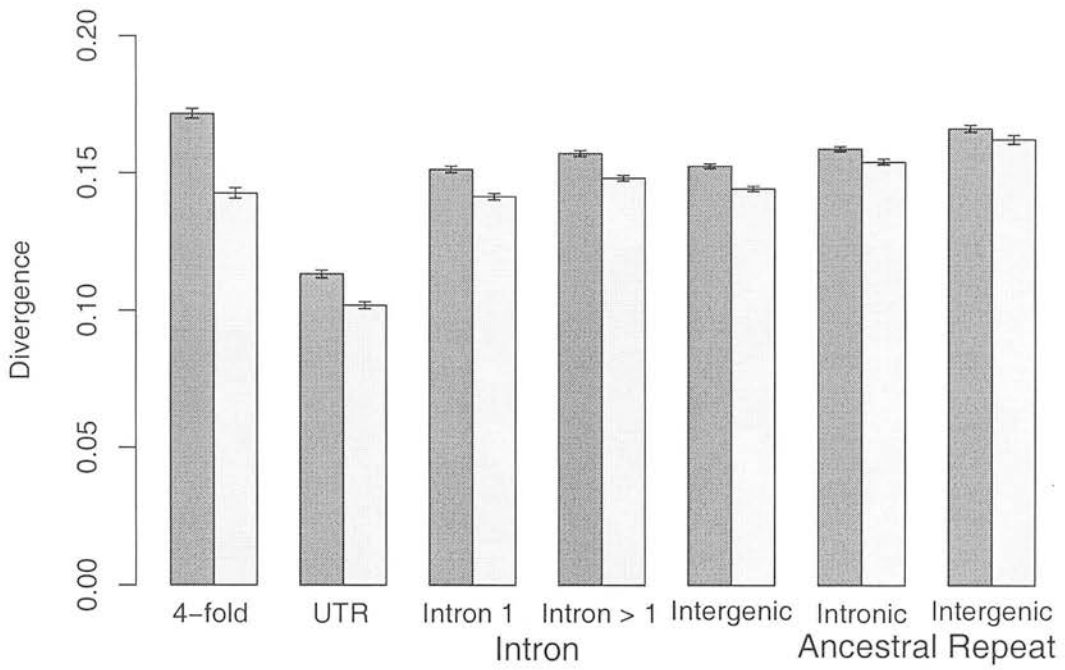

Figure 5.2.: Mean nucleotide substitution rates for different sequence types. Dark bars show substitution rates estimated at all sites, light bars show rates at non CpG-prone sites. Intronic substitution rates were estimated excluding splice regions which were assumed to occur in the first 20 and last 40 base pairs. 95% confidence intervals were estimated by bootstrapping the dataset by 1Mb block.

classes, with the widest margin between the two observed at fourfold degenerate sites and the smallest within intergenic repetitive DNA. This gradient reflects the variation in the CpG content of each sequence class. At non CpG-prone sites, the most swiftly evolving sequence class are ancestral repeats. This supports the assumption that, after excluding CpG dinucleotides, transposable elements contain the highest proportion of neutrally evolving sites. The results also show a significantly higher (5% at non CpG-prone sites) nucleotide substitution rate in those transposable elements located within intergenic DNA. This is evidence for a lower base mutation rate in transcribed DNA, possibly due to transcription coupled repair of genes expressed in the germline (Hanawalt, 1994). The non CpG-prone nucleotide substitution rate estimated in intronic transposable elements (15.4%) is very similar to the mean non CpG-prone substitution rate estimated in the assumed neutral standard (the fastest evolving intronic sites) in a previous study (15.6%; Keightley and Gaffney (2003)).

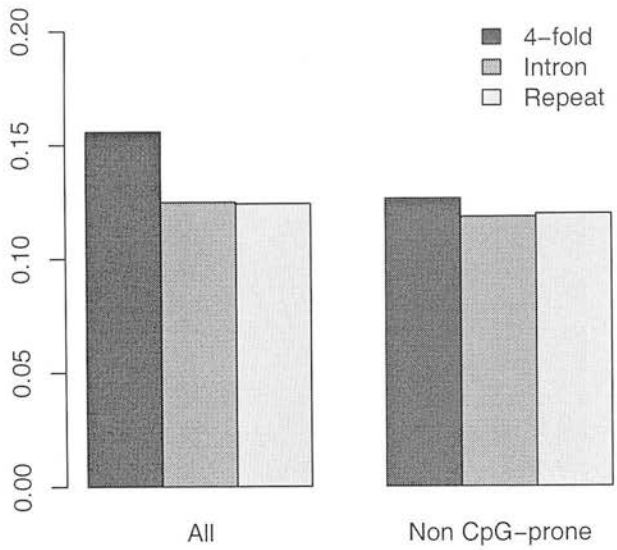

Figure 5.3.: Estimated substitution rates at simulated fourfold degenerate, intronic and repetitive sites. Each estimate is the mean over 100 simulated replicates, each of which evolved a single sequence containing $\sim$ 8Mb of coding, intronic and repetitive sequence along two lineages.

In order to investigate the effects of base composition and site selection on estimates of nucleotide substitution rates, real fourfold sites, introns and ancestral repeats were simulated to evolve down two lineages entirely free of selective constraints. Mean nucleotide substitution rates of each of the simulated sequence classes across 100 simulated phylogenies is presented in Figure 5.3. At all sites, substitution rates are most substantially affected by differential frequencies of the CpG dinucleotide and are, therefore, highest at fourfold degenerate sites, which have the highest CpG frequency (0.044, compared with 0.010 and 0.008 in introns and ancestral repeats, respectively). In addition to CpG hypermutation, the mutation model incorporates marginally more mutable GCs than ATs ($\sim$ 1.36-fold). It is presumably this which produces the decreasing gradient in non CpG-prone substitution rates from fourfold degenerate sites (%GC = 57.4), ancestral repeats (%GC = 44.5) to introns (%GC = 43.0). Compared with the rates of nucleotide substitution observed in the real sequence data, a number of patterns are evident. Firstly, although fourfold sites are

clearly the most swiftly evolving class in the neutral simulations, this is not the case at non CpG-prone sites in the real data (Figure 5.2). Whilst simulated fourfold sites evolved $\sim 7\%$ faster than repetitive DNA, real fourfold sites are in fact evolving $\sim 7\%$ slower than real intronic transposable elements. Thus the real data suggest a low level of purifying selection at murid fourfold degenerate sites.

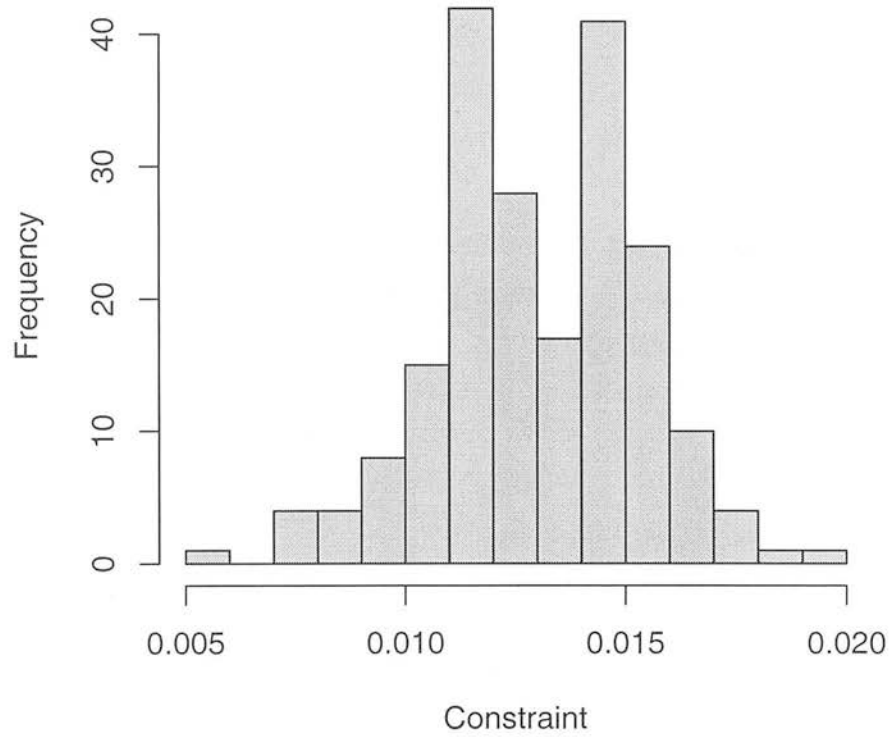

Figure 5.4.: Distribution of "constraint" values estimated in neutrally evolved, simulated introns, using transposable elements evolved under the same mutational model as a "neutral standard". Constraint was estimated for each one of 100 replicates.

Secondly, in the simulated phylogenies, transposable elements are evolving marginally faster ($\sim 1.3\%$) than introns. This is also the case in the real data and suggests that at least some of the elevated evolutionary rate observed in repetitive sequence over intronic DNA is due to neutral, mutational effects coupled with slight compositional variation. Constraint was also estimated in simulated

intronic sequence. For each simulated replicate constraint was estimated in an identical fashion to the real data. Simulated transposable elements were used as a neutral standard to calculate the expected numbers of substitutions in the simulated intronic sequences, and this was compared to the observed rate. The distribution of estimated constraint across 100 replicates (Figure 5.4) suggests that positive constraint values, up to a maximum of ~2%, could be explained by mutation/compositional bias alone. Although the smallest difference observed between repetitive and intronic evolutionary rates ($\sim$ 3.8%; between intronic transposable elements and non first introns, excluding splice regions) is still larger than that observed in simulated data, this result suggests caution in the interpretation of differences in substitution rate between sequences of even marginally different base composition.

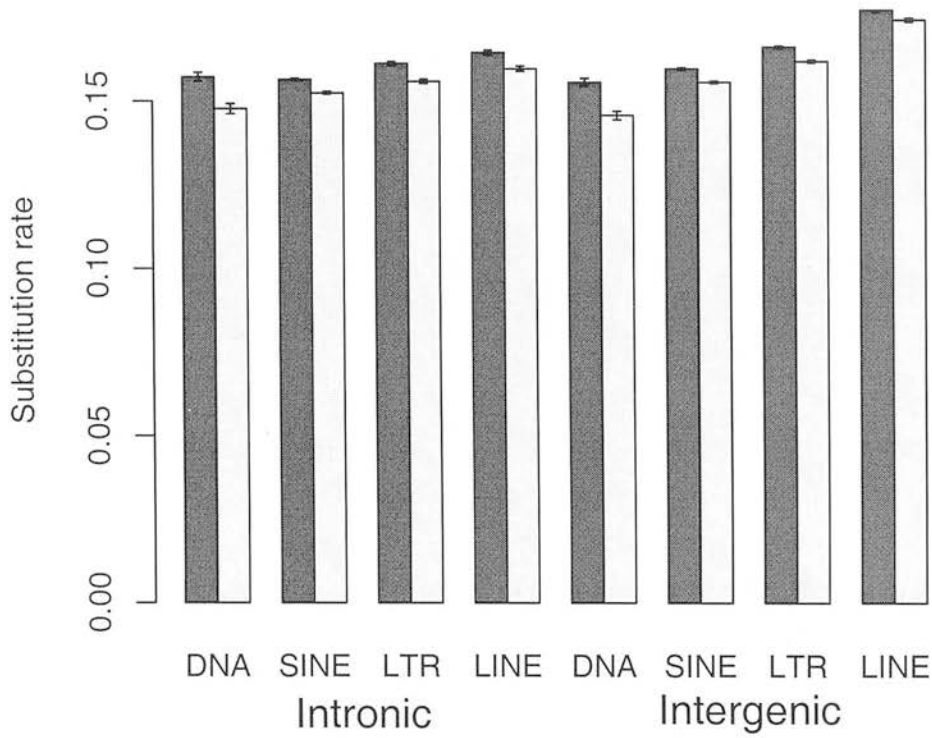

Figure 5.5.: Estimated mean nucleotide substitution rate across the four transposable element families used as putative neutral standard: SINEs, LINEs, LTRs and DNA. Elements are subdivided into those found in intronic and intergenic sequence. 95% confidence intervals were estimated by bootstrapping the dataset by 1Mb block.

Variation in Ancestral Repeat substitution rates

It was necessary to investigate the validity of the assumption of neutral evolution in ancestral repetitive DNA. As part of this the mean nucleotide substitution rate was estimated separately in each of the four main transposable element classes: Short INterspersed Elements (SINEs), Long INterspersed Elements (LINEs), Long Terminal Repeat retrotransposons (LTRs) and DNA transposons (DNA). The results of this analysis are presented in Figure 5.5. Under the assumption of selective neutrality and the same base mutation rate, the nucleotide substitution rates in each element class should be approximately the same. It is evident, however, that this is not the case and there is a clearly significant variation in the mean element family substitution rates. This pattern is evident in both intronic and intergenic elements, although intergenic elements tend to be slightly more swiftly evolving than their intronic counterparts.

X vs Autosome Substitution rates

Substitution rates between the autosomes and the X chromosome were also compared (Figure 5.6). The results show that the proportional difference between substitution rates at all sites and non CpG-prone sites is reduced on the X chromosome. This perhaps suggests that there is a reduced level of CpG hypermutability on the X chromosome, when compared with the autosomes. This disparity is most evident at fourfold degenerate sites, where CpG frequencies are highest. Here, the difference between substitution rates at all sites and non CpG-prone sites is substantially reduced on the X chromosome. The effect is smaller within transposable elements, which are the most CpG poor of the sequences in this study. This result contradicts the findings of another recent study which suggested that male mutation bias is stronger at non CpG sites than at CpG sites (Taylor et al., 2006). If this were the case in the current dataset a larger proportional difference in substitution rate between all and non CpG-prone sites would be expected on the X chromosome. In fact, estimates of the male-to-female mutation bias in this study are larger when all sites are considered than rather than non CpG-prone sites (e.g. $\alpha \sim 4.78$ at all fourfold sites, $\alpha \sim 2.40$ at non CpG prone fourfold sites).

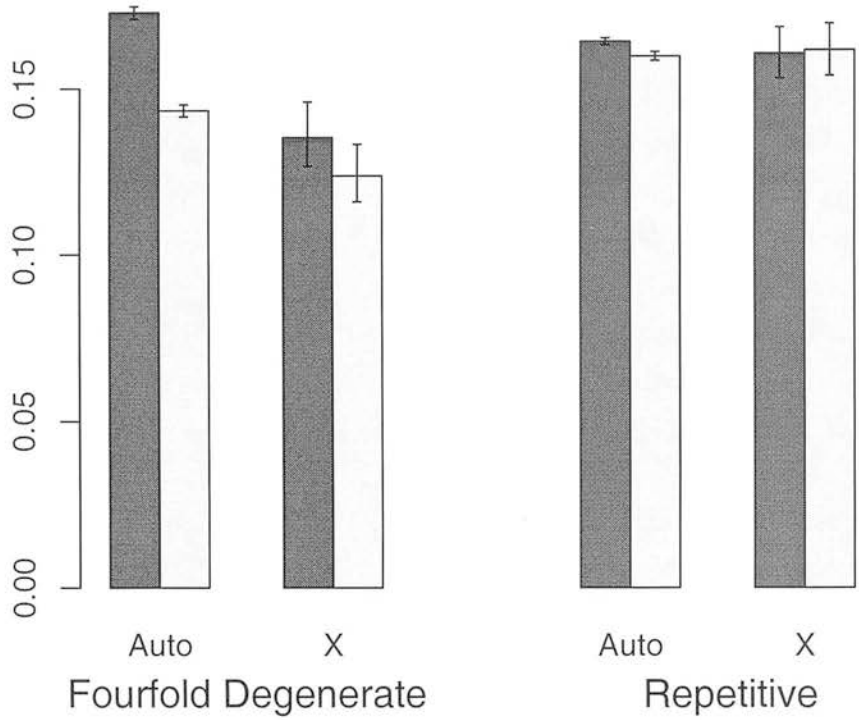

Figure 5.6.: Repetitive and fourfold degenerate nucleotide substitution rates at all and non CpG-prone sites on the autosomes and X chromosome. Here repetitive sequence refers to both intronic and intergenic transposable elements. 95% confidence intervals are shown and were estimated by bootstrapping the data by 1Mb block, 1000 times.

Divergence and Constraint

Variation in divergence and constraint with distance from splice sites in first introns is shown in Figures 5.7 (a) and 5.8. It is clear that, whilst divergence is lowest at the 5′ end of intron 1, it remains below the average intronic ancestral repeat substitution rate for a considerable distance from both 5′ and 3′ splice sites. This is clearly reflected in the level of constraint estimated in intron 1 (Figure 5.8). Constraint is significantly above zero in first introns for at least the first 10kb upstream and downstream of the acceptor and donor splice sites, respectively. As demonstrated previously (Keightley and Gaffney, 2003) constraint is highest at the 5′ end of intron 1, reaching a maximum value of approximately 20% immediately adjacent to the 5′ splice site. In contrast to

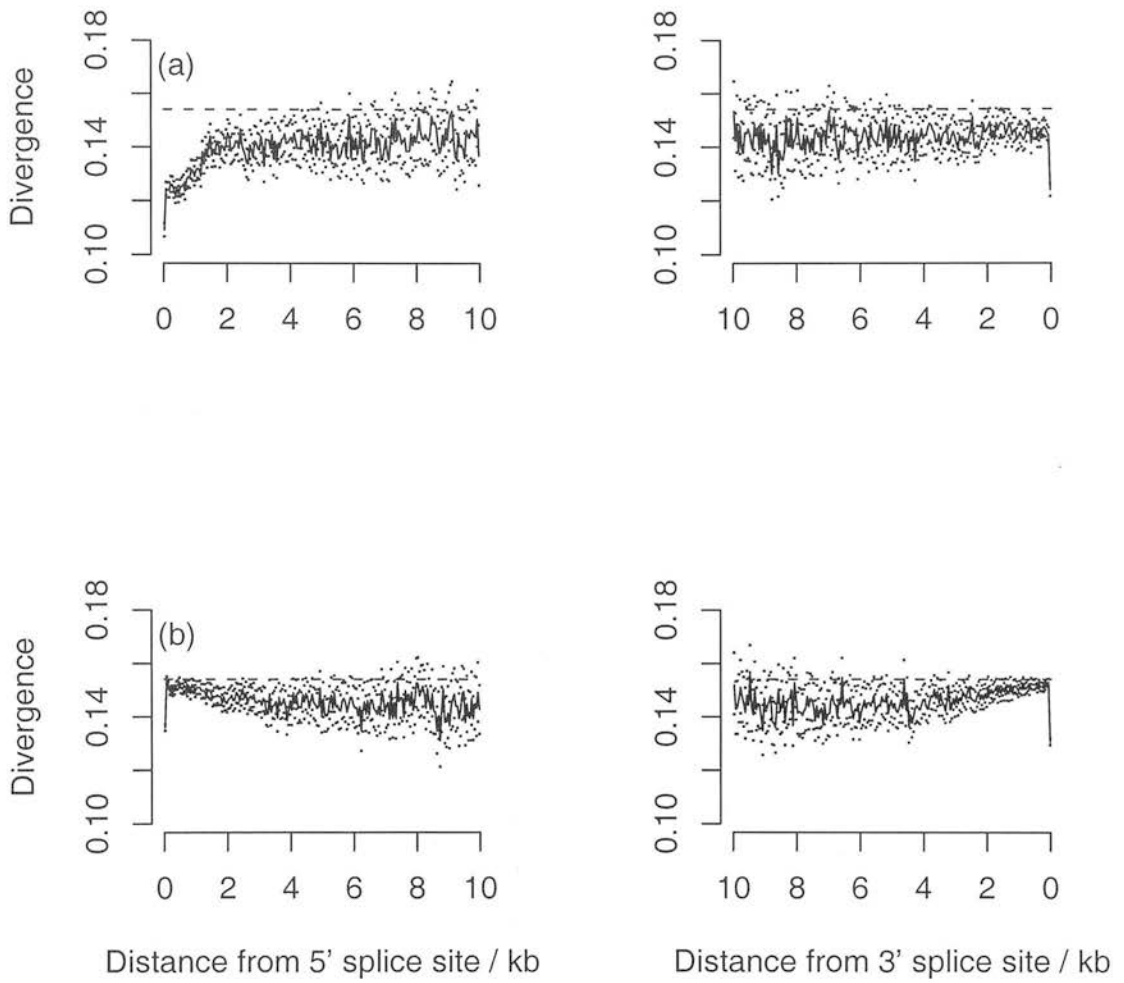

Figure 5.7.: Pairwise divergence at non CpG-prone sites in 50bp blocks in intron 1 (a) and non-first introns (b), upstream and downstream of 3′ and 5′ splice regions. Dots show the 95% confidence intervals for each block substitution rate and were estimated by bootstrapping the data by 1Mb block, 1000 times. The dotted line shows the mean divergence of intronic transposable elements.

this previous analyses, however, adoption of a new neutral standard and a larger dataset reveals that the 3′ end of intron 1 is also under low but significant, constraint.

Nucleotide substitution rates in non first introns were also estimated (Figure 5.7 b). It appears that pairwise divergence in non-first introns decreases with distance from the intron-exon boundary, becoming significantly different from repetitive element divergence at a distance of approximately 3kb from both the

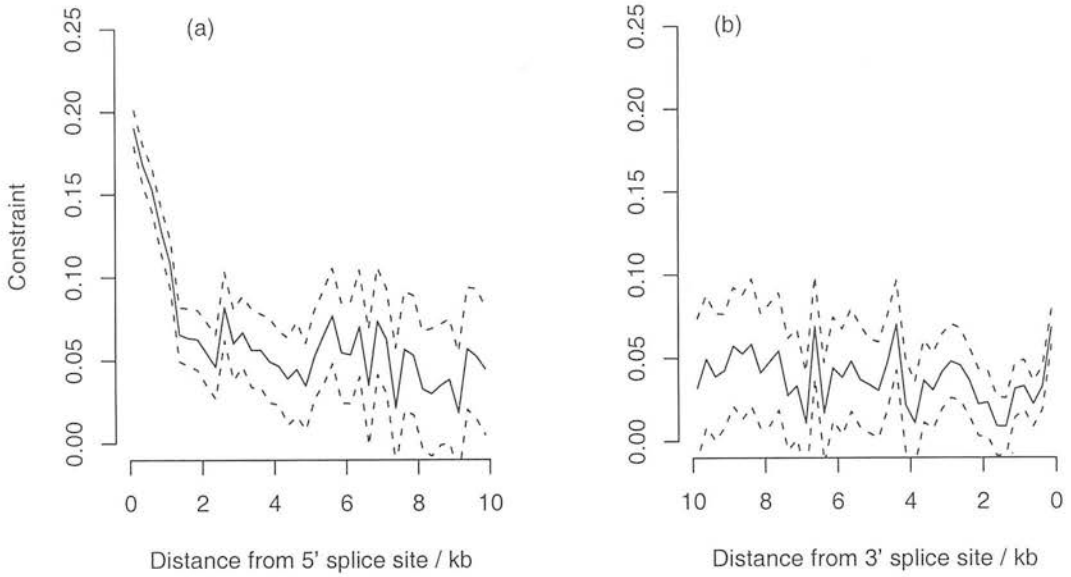

Figure 5.8.: Change in constraint in intron 1 with distance downstream (a) and upstream (b) of the donor and acceptor sites, respectively. Dots show 95% confidence intervals estimated bootstrapping the data by 1Mb block.

5′ and 3′ splice sites. However, this does not reflect variation in divergence within all non-first introns but rather a weak negative correlation between intron length and divergence because, by definition, only introns >3kb in size contributed to estimates of divergence beyond this point in the plots.

In our previous analyses (Keightley and Gaffney, 2003) it was assumed that the fastest evolving intron sites were neutrally evolving. This assumption, by definition, precluded the estimation of selective constraint in those introns in which the neutral standard was located. The use of ancestral repeats in the current study allowed, for the first time, the investigation of patterns of selective constraint in non first introns (Figure 5.9). The results show that, whilst intronic sequence situated in the first 1-2kb of non first introns is evolving neutrally (assuming intronic transposable elements are also evolving neutrally) sequence more distant than this from the splice sites is under a low to moderate selective constraint. Although this may seem counterintuitive, it again results from the relationship between non first intron length and divergence/constraint (Figure 5.12). Thus, whilst introns under ~6kb are evolving at the same rate as intronic transposable elements (Table 5.2), introns longer than this appear to be selectively

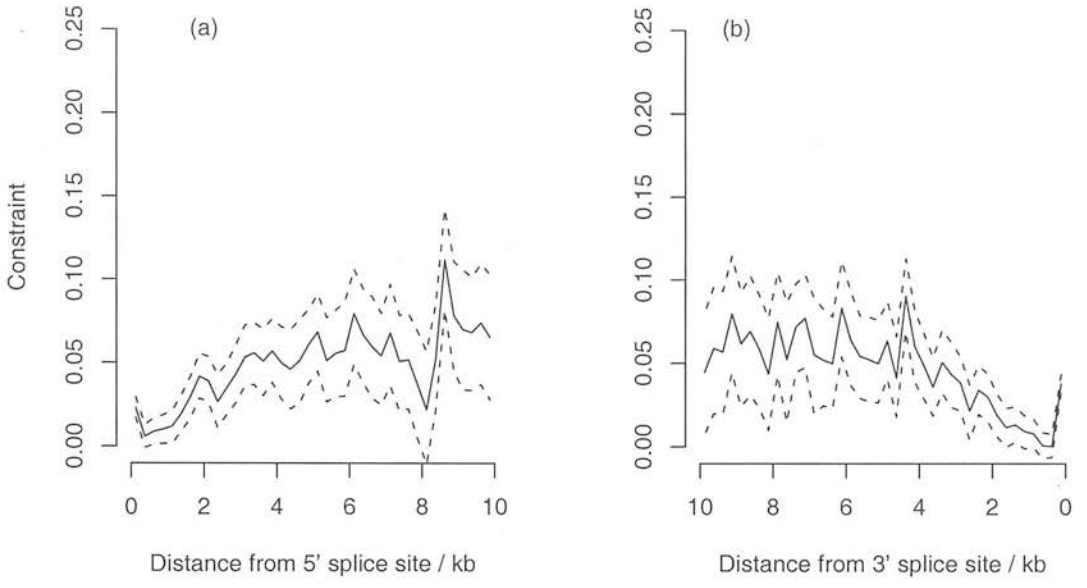

Figure 5.9.: Change in constraint in non first introns with distance downstream (a) and upstream (b) of the donor and acceptor sites, respectively. Dots show 95% confidence intervals estimated bootstrapping the data by 1Mb block.

constrained.

The change in mean pairwise divergence (Figure 5.10) and constraint (Figure 5.11) in intergenic DNA with distance from the transcription start/stop points was also estimated. Although there is a sharp drop in constraint immediately adjacent to the start/end of the UTRs, this appears to plateau at $\sim$ 5kb. Further into the intergenic region, constraint apparently does not drop to zero but appears to increase slightly. Although the number of sites also decreases with distance it seems that, even comparatively large distances from genic regions, alignable, nonrepetitive sites are still under moderate selective constraint. It is notable that the 95% confidence intervals of constraint estimates do not span zero at any point over a distance of 50kb (Figure 5.11).

In order to calculate the relative contribution of each different sequence class to the total numbers of constrained sites in the genome the mean selective constraint across all sites for each class was estimated. The number of constrained bases per locus was defined as the product of the mean constraint and the mean number of aligned sites per locus for that class. To get the total number of sites, this figure was multiplied by an estimate of the total number of mouse genes. The estimate

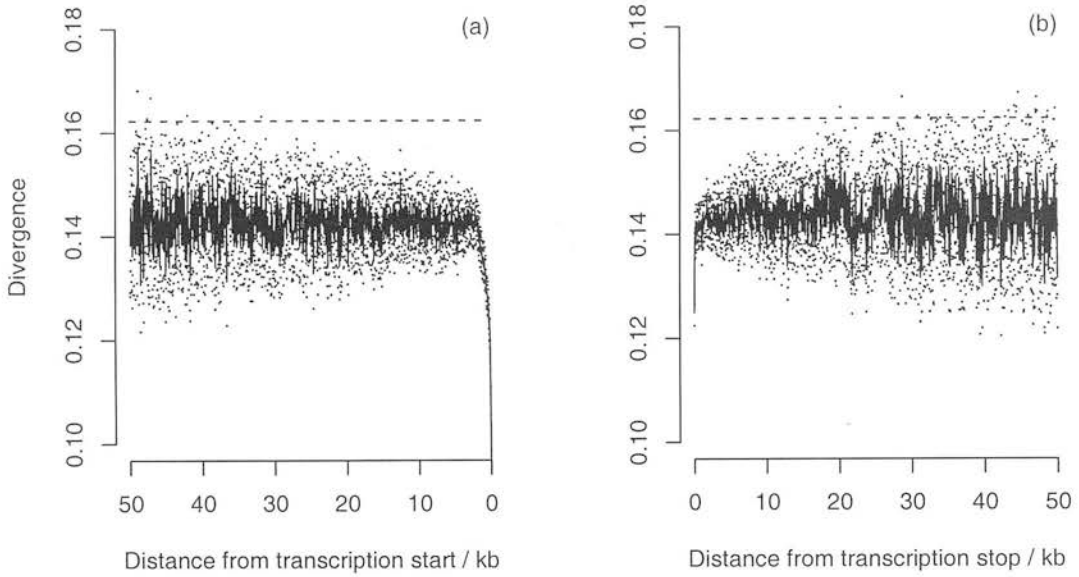

Figure 5.10.: Pairwise divergence at non CpG-prone sites in 50bp blocks in intergenic DNA upstream (a) and downstream (b) of annotated transcription start points. Dots show the 95% confidence intervals for each block substitution rate and were estimated by bootstrapping the data by 1Mb block, 1000 times. The dotted line shows the mean divergence of intergenic transposable elements.

of the number of mouse genes (26512) was based on the total number of known and predicted genes in release 36 of the ENSEMBL database (Hubbard et al., 2005). These estimated contributions are presented in Table 5.2. A few striking patterns are evident. Firstly, whilst the estimated number of constrained, nondegenerate coding sites is not insubstantial (25 Mb), there are over 3 times as many constrained sites in noncoding regions (83 Mb). In addition, of all classes of noncoding DNA, the majority (47 Mb) of constrained sites are located within the "deep" (>5kb from known coding sequence) intergenic regions. The contribution of intronic sequence to the total number of constrained bases is, by comparison, small. However, these results would suggest that only short (< 6kb) introns are evolving at approximately the same rate as intronic ancestral repeats.

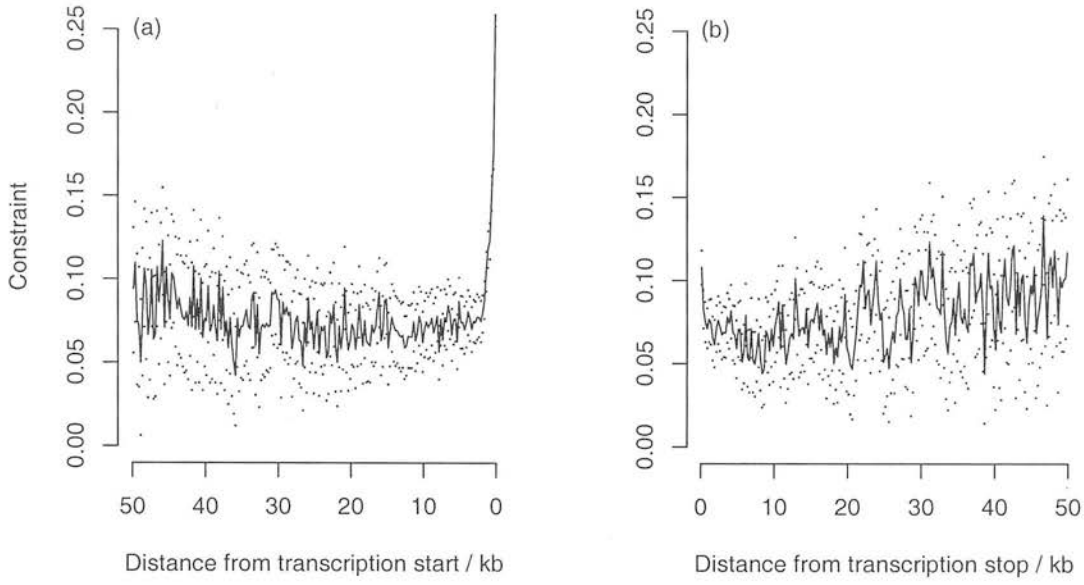

Figure 5.11.: Change in constraint intergenic DNA with distance upstream (a) and downstream (b) of transcription start and stop points, respectively. Dots show 95% confidence intervals estimated by bootstrapping the data by 1Mb block, 1000 times.

Intron length divergence correlation

Previous results in this study suggest that there is a relationship between intron length and divergence/constraint. A plot of mean divergence versus intron length is shown in Figure 5.12 and reveals a weak correlation between non-first intron length and intronic substitution rate. There is a striking contrast between first and non-first introns. Even with splice sites removed, a considerably lower divergence is observed in short (<6kb) first introns. In comparison, most short non-first introns are evolving at the same rate as intronic transposable elements. It also seems that, apart from highly constrained short first introns there is little or no relationship between length and divergence in first introns. Similar relationships between intron length and divergence have recently been reported in *Drosophila* (Haddrill et al., 2005). The relationship observed here is, however, considerably weaker than that seen in fruitflies.

5.4. Discussion

This study has investigated genomic patterns of selective constraint in murids. These results provide comprehensive estimates of genomic selective constraint

Table 5.2.: Mean selective constraint across all aligned sites for different classes of noncoding DNA and fourfold degenerate sites. Intron splice sites were defined as the first 20bp downstream of the donor splice site and last 40bp upstream of the acceptor splice site. For estimates of constraint within non first introns, "$\leq, >$ 6kb" refers to estimates of constraint within introns in which the *total* intron length in mouse is less than or greater than 6kb. For intergenic DNA, "$\leq, >$ 5kb" refers to estimated constraint *within or beyond* the first 5kb upstream/downstream of transcription start or stop point. Number of sites per locus refers to the mean number of aligned sites per locus for each sequence type. Number of constrained bases is an estimate of the number of bases in the alignable portion of the genome that are completely constrained. Estimates of the total number of constrained bases were calculated assuming 26512 mouse genes. This is the total number of known and predicted mouse genes in release 36 of the ENSEMBL online database (Hubbard et al., 2005). 95% confidence intervals for each estimate are shown in parentheses.

Sequence Type		Constraint	bp/Locus	Constrained Bases
CODING		0.87 (0.89,0.85)	1125	25.07 Mb †
FOURFOLD		0.044 (0.057,0.032)	198	231 kb
UTR		0.319 (0.328,0.311)	609	4.97 Mb
5′		0.470(0.483,0.455)	91	1.09 Mb
3′		0.294 (0.303,0.285)	517	3.90 Mb
INTRON (splice sites)		0.215 (0.221,0.209)	380	2.17 Mb
Intron 1		0.264 (0.281,0.248)	48	336 kb
Intron $\neq$ 1		0.205 (0.212,0.198)	332	1.8 Mb
INTRON (excl. splice)		0.050 (0.056,0.045)	13075	17.3 Mb
Intron 1		0.080 (0.087,0.074)	4065	8.62 Mb
	all	0.036 (0.042,0.029)	9009	8.60 Mb
Intron $\neq$ 1	$\leq$ 6 kb	-0.003 (0.002,-0.008)	4699	
	$>$ 6 kb	0.071 (0.079,0.063)	4310	
INTERGENIC		0.108 (0.113,0.104)	20958	58.97 Mb
5′	$\leq$ 5 kb	0.122 (0.127,0.117)	2240	7.00 Mb
	$>$ 5 kb	0.103 (0.108,0.098)	9032	23.83 Mb
3′	$\leq$ 5 kb	0.086 (0.092,0.080)	1884	4.15 Mb
	$>$ 5 kb	0.115 (0.122,0.108)	7801	22.98 Mb

†(Keightley and Gaffney, 2003)

in a variety of classes of murid noncoding DNA. The findings of this study can be summarised as follows. Firstly, the data strongly suggest low but extensive selective constraint within intergenic DNA. Importantly, this does not

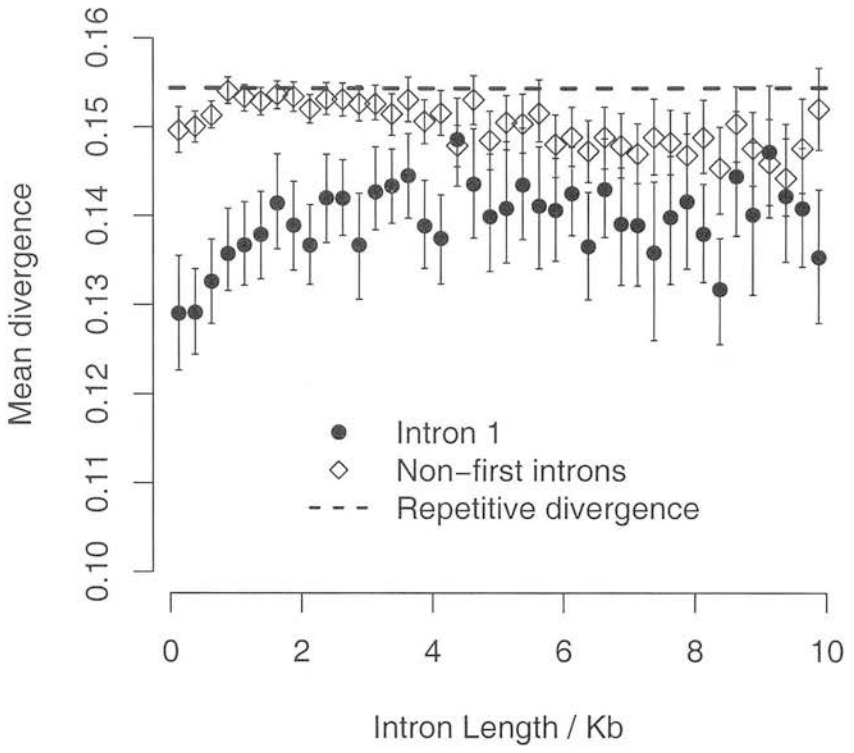

Figure 5.12.: The relationship between intron length and non CpG-prone divergence for first and non-first introns. Introns were divided by their total length in mouse into classes and the mean divergence for each class estimated. Splice sites were excluded from estimates of divergence. 95% confidence intervals were estimated bootstrapping the data by 1Mb block, within length class, 1000 times. The dotted line shows mean ancestral repeat non CpG-prone divergence.

appear to decrease with distance from known genic regions beyond a certain distance (~5kb). Secondly, this study clarifies the extent of evolutionary constraints within introns. As previously suggested (Majewski and Ott, 2002; Keightley and Gaffney, 2003; Chamary and Hurst, 2004, 2005), murid first introns appear to harbour proportionally more evolutionarily constrained sequence than non first introns. However, this analysis also shows that low constraint in non first introns also exists. Intriguingly the level of constraint in non-first introns is apparently positively correlated with intron length. Assuming that nucleotide substitution rates within transposable elements primarily reflect mutational rather than selective forces, these results also provide some insight into patterns of mutational variation, namely that (i) mutation rates within

transcribed DNA are lower than those in nontranscribed regions and (ii) the level of CpG hypermutability may be marginally reduced on the X chromosome, compared with the autosomes. Finally, these results suggest that there are some differences in the mean substitution rates between different families of transposable elements.

The agreement of these results with our previous analysis (Keightley and Gaffney 2003; Chapter 4) is discussed first. Assuming that the majority of transposable elements in the current dataset are evolving neutrally, these data suggest that the assumption of neutral evolution for fastest evolving sites of non first introns is a reasonable one. The rate of nucleotide substitution estimated previously at the fastest evolving intron sites compares well with the rates estimated at intronic ancestral repeats. A major departure of the current work from our earlier study (Keightley and Gaffney, 2003) is the evidence of significant selective constraint within intergenic sequence at distances greater than 3-4kb from a coding sequence. There are two reasons for this. Rates of nucleotide substitution within intergenic and intronic transposable elements suggest a marginally, but significantly, lower mutation rate within transcribed DNA, possibly the result of transcription-coupled-repair. Thus, it seems that fastest evolving intron sites are an inadequate model for neutral evolution in intergenic DNA. Using such sites as a neutral standard means that the expected change (and therefore total constraint) in intergenic DNA was underestimated. Perhaps more importantly, in the current study all transposable elements in noncoding DNA were excluded from the estimation of observed rates of nucleotide substitution. Given that repeats are the most swiftly evolving class of noncoding sequence in this study, it is likely that the inclusion of repetitive sequence substantially reduced the estimate of intergenic constraint in the previous study. A similar change in constraint as observed in the previous study with distance upstream and downstream from the coding sequence is observed here. Estimates of constraint immediately adjacent to coding sequence are significantly lower than in than the previous study, due to the treatment of gene UTRs as a separate class.

This study also provides an insight into the evolutionary significance of introns. Perhaps surprisingly, introns appear to be the least constrained class of noncoding DNA in murids. The most highly constrained intronic sequence is located within intron 1. This pattern would suggest that motifs involved in gene regulation and splicing tend to be preferentially located in intron 1. In this dataset first introns

tend are over three times as large as non first introns (8159bp and 2656bp, respectively). This size differential also likely reflects the preferential situating of functional motifs within the first intron. It is notable, however, that even in intron 1 constraint is lower than in either nearby or distant intergenic sequence.

Outside intron 1, constraints tend to be lower. Constraint in non first introns is related to total intron length, such that introns less than 6kb in length appear to be evolving at approximately the same rate as intronic transposable elements, whilst introns larger than this are constrained at a level approaching that found in intron 1. This pattern has previously been shown to exist in *Drosophila* (Haddrill et al., 2005). In addition, in humans there is evidence that multispecies conserved sequences are preferentially located within longer introns (Sironi et al., 2005). This study presents a quantitative assessment of this phenomenon in murids. An immediate explanation for the relationship between selective constraint and length is lacking. However some speculative hypotheses can be suggested. These speculations rely crucially on the relationship between intron size and age. It is likely that, since their origin, mammalian genomes have increased in size. This may have occurred due to invasion by multiple, highly-replicating families of transposable elements combined with a reduction in the efficacy of natural selection, as a result of decreasing effective population sizes, to remove such "junk" sequences. In addition, there is a positive correlation between mammalian intron and genome size (Vinogradov, 1999) and it therefore seems likely that there is also a positive correlation between intron length and age. If we accept that this is the case then there are at least two scenarios whereby longer introns could contain proportionally more functional sequence. It has been suggested that intron presence alone can be advantageous in a coding sequence, as a means of detecting aberrant mRNA molecules, via nonsense-mediated decay (Lynch and Kewalramani, 2003). If it is via this selective advantage that introns initially become established it would suggest that smaller, younger introns in a coding sequence function primarily in this capacity whilst any functional motifs will tend to be preferentially located in larger, older introns. Secondly if mammalian introns expand with age and, over the course of evolutionary time, motifs within introns are co-opted for function then the longer/older introns will harbour more selectively constrained, functional DNA. Notably, however, that it is only if the rate of co-option of sequence for function outstrips the rate of intron expansion that the length/constraint correlation observed here could be produced.

These results are relevant to the study of CNGs. It is likely at least some of the signal of selective constraint, both within introns and intergenic DNA, is derived from the multispecies conserved regions that have recently been subject of much scrutiny (Thomas et al., 2003; Dermitzakis et al., 2005; Siepel et al., 2005). This study enables a quantitative comparison of the evolutionary significance of such sequences with regions with a more clear cut biological function, such as coding sequence. Perhaps surprisingly, the results suggest that the majority of sites preserved by purifying selection are located at distances greater than 5kb from genic regions. Furthermore, results presented here indicate that, in the murid genome, the quantity of selectively constrained sequence in these "deep" intergenic regions alone is twofold greater than in nondegenerate coding sequence. The analysis suggests that a total of 76.27 Mb of constrained, functional DNA is located either within intergenic regions or outside the splice sites of introns. Given that the quantity of constrained sites located within CNGs has been estimated at approximately 49Mb (Keightley et al., 2005a), this suggests that a substantial amount of functional sequence remains to be discovered in murids.

This study adds to the growing body of work which indicates that mammalian synonymous sites are under some purifying selection (Parmley et al., 2006; Chamary and Hurst, 2005, 2004). However, this dataset show that, in murids at least, the selective constraint at such sites is extremely weak ($\sim$ 37% less than estimated in long, non-first introns). The GC content at fourfold sites is elevated beyond that typical in noncoding DNA ($\sim$53%) and this may reflect codon bias towards G- or C-ending codons, as has been reported in humans (Kondrashov et al., submitted). Selective constraint at fourfold sites could result from conserved motifs, such as exonic splice enhancers (Fairbrother et al., 2004; Parmley et al., 2006).

These results also shed light on two potential aspects of mutational variation. Firstly, substitution rates within transposable elements strongly suggest some effect of transcription coupled repair. An explicit test of this hypothesis would be to compare nucleotide substitution rates of genes that are expressed in the germline versus those that are not. Secondly, these data suggest a weak reduction of the level of CpG hypermutability on the murid X chromosome. As mentioned, this results contradicts the findings of a recent comparison of human-chimpanzee sex chromosomes and autosomal substitution rates (Taylor et al., 2006). There is no clear explanation for these differences. The results of the two studies are

not directly comparable, however, as Taylor et al. (2006) divide sites into CpG and non CpG, as opposed to non CpG prone. In addition, it is known that male mutation bias murids is much smaller than in hominids (IRGSC, 2004), presumably as a result of differences in generation time. It is possible that reduced CpG hypermutability on the X chromosome may be a selective response to reduce exposure of recessive deleterious mutations in the hemizygous state (McVean and Hurst, 1997). However, further work is required to resolve this conflict.

The accuracy of estimates of constraint relies heavily on the assumption of neutral evolution in transposable elements. However, significant variation in mean nucleotide substitution rate between different repeat families indicates that this may not be the case. Recent work has suggested that primate repetitive DNA is more heavily methylated than nonrepetitive (Meunier et al., 2005), and the rate of transitions at CpG dinucleotides is correspondingly elevated. However, even if this also occurs in murids, it should have relatively little impact upon estimates of constraint, all of which were calculated excluding CpG-prone sites, an efficient method of removing most CpG mutations (*see* Chapter 2). Other work has also indicated that some repeats occasionally acquire a selectively beneficial function (Britten, 1997; Kamal et al., 2006) and thus be preserved by purifying selection. Alternatively, differences in substitution rate may be driven by compositional variation, although there is no clear relationship between nucleotide substitution rate and either GC or CpG content. Another hypothesis is that different families of repeats insert into regions with different mean mutation rate. It is known that SINE elements are preferentially located within GC rich regions in the human genome, whilst LINE elements tend to occur in AT rich regions (IHGSC, 2001). If such regions correspond to regions with different mutational regimes, then this could explain the differences in substitution rates between transposable element families. Finally, it may be that substitution rates in transposable elements are influenced by biased gene conversion (Webster et al., 2005) such that mutations increasing GC content are fixed preferentially. The effect of biased gene conversion (BGC) is equivalent to selection for the allele towards which gene conversion is biased (Nagylaki, 1983). If this process is occurring then this could explain some of the differences in substitution rate between transposable element families observed.

If, for any of these reasons, the transposable elements used here as a neutral

standard throughout this study are evolving non-neutrally then estimates of constraint will be incorrect. In particular, if substantial selection or BGC is occurring, it is likely that constraint in murid noncoding DNA has been underestimated. However, in this study it appears that repetitive DNA is the most swiftly evolving sequence class. Thus, if such processes are indeed occurring in transposable elements their effects are small and/or confined to a minority of elements within the dataset. A further caveat relates to the exclusion of potentially alternatively spliced genes from the dataset. It is known that introns within alternatively spliced genes are more highly conserved than in constitutively spliced genes (Sorek and Ast, 2003). If such conserved regions do function in the regulation of alternative splicing, then estimates of intronic constraint may be biased downwards.

6. Discussion

Noncoding DNA makes up the greatest proportion of many eukaryotic genomes. Despite this, we know relatively little about the possible functions and evolutionary significance of most noncoding sequence. The extent of functional noncoding DNA is relevant to the study of genetic disease, as well as being important in evolutionary theory. In this thesis, the extent and genomic location of selectively constrained noncoding DNA in murids has been investigated. In addition, aspects of mutational variation and bias, relevant to the study of the evolution of noncoding DNA, have been addressed.

In Chapter 2 the assignment of ancestral CpG state at fourfold degenerate and noncoding sites was addressed. Because CpG dinucleotides are hypermutable in mammals, many molecular evolutionary studies attempt to separate nucleotide substitution rates into those which occur within and outside a CpG dinucleotide. The efficacy of ancestral CpG assignment was assessed using simulations of a simple, two-branch phylogeny. Results of these simulations show that assignment of the ancestral CpG state based in the presence/absence of the CpG dinucleotide in two derived lineages can lead to seriously biased estimates of the nucleotide substitution rate at both CpG and non CpG sites. The reason for the inaccuracy of this assignment method is the large variation in the numbers of true CpG changes that are misassigned as non CpG and *vice versa*. Specifically, across small evolutionary distances, substantially more non CpG changes are misassigned as CpG whilst across large evolutionary distances, the opposite is the case. The results presented in this chapter also show that the level of misassignment depends crucially on the base composition of the ancestral sequence and, to a lesser extent, on the degree of CpG hypermutability. The effects of misassignment biases at fourfold degenerate and noncoding sites was also addressed. It was shown that, because of the structure of the genetic code, fourfold degenerate sites in coding sequence tend towards a much higher CpG frequency than noncoding sites. Because of these compositional differences, the impact of CpG misassignment is different at fourfold degenerate and noncoding sites. Specifically, simulations showed that across small evolutionary distances, both the CpG and non CpG nucleotide substitution rates estimated at fourfold

degenerate sites are lower than those estimated in noncoding sequence, despite having been evolved under identical mutational regimes and without any selection. Thus, fourfold degenerate sites can appear to be evolving more slowly than noncoding sites as a result of problems with ancestral CpG assignment. This result is relevant to comparisons of species which are relatively closely related and whose genomes are compositionally similar to murids. In particular, these results have a bearing on recent work which has compared rates of nucleotide substitution at human and chimp synonymous and noncoding sites (Hellmann et al., 2003; ICGSC, 2005). These studies have indeed inferred purifying selection at synonymous sites, from the observation that rates of nucleotide substitution at CpG and non CpG sites, estimated using CpG/ non CpG assignment, are 30-50% lower at synonymous sites than in noncoding DNA. The results presented in this chapter suggest that at least some of this difference in substitution rates is due to misassignment biases.

In Chapter 3 the scale and magnitude of mutational variation in murids was investigated. There is good evidence to suggest that the mutation rate varies considerably in mammals (Wolfe et al., 1989; Ellegren et al., 2003). Estimating the scale of this variation is relevant to attempts to locate selectively constrained, putatively functional regions via comparative methods, since the mutation rate determines the null hypothesis that sequences will be conserved by chance. In addition, patterns of mutational variation can improve our understanding of the processes that drive mutation. The scale and level of mutational variation were inferred in this chapter using estimated nucleotide substitution rates in transposable elements, under the assumption that transposable elements are evolving neutrally. The results presented show that the primary scale over which point mutation rates vary is approximately 1Mb. This result is supported by two lines of evidence. Firstly, plots of the partial autocorrelation coefficient suggest that similarity of mutation rates beyond a distance of 1Mb can be explained by the "propagation" of lower order autocorrelation across distances of up to 1Mb. Secondly, when the data were fitted to a mixed model which included terms for inter- and intra-chromosomal effects of different sizes, the most parsimonious model included a term for a 1Mb regional or intra-chromosomal random effect. The results of the mixed model analysis also showed that mutational variation along the length of a chromosome is substantially greater than that between chromosomes.

Chapters 4 and 5 concern the magnitude of, and spatial variation in, selective constraint within murid noncoding DNA. In Chapter 4 a moderately sized dataset of 300 mouse-rat gene orthologues was used to compare the rate of nucleotide substitution in intronic and intergenic noncoding DNA. In order to estimate selective constraint within these regions it was assumed that certain sites within introns were evolving neutrally. The rationale was to identify those intronic sites which were the most swiftly evolving, under the assumption that this high evolutionary rate was unlikely to be due to substantial adaptive evolution within introns. The fastest evolving sites in introns were located within non-first introns, outside the regions presumably involved in splicing. Using these sites as a "neutral standard" the level of constraint within intergenic DNA was shown to be substantial and it was estimated that approximately one in every three new mutations occurring immediately upstream (downstream) of the start (stop) codons in murids was strongly deleterious and removed by purifying selection. The level of selective constraint in intergenic DNA was estimated to remain significantly above zero until approximately 4kb upstream and downstream of the coding sequence. Beyond this distance, constraint was statistically indistinguishable from zero, on average. Selective constraint in the $5'$ region of the first intron was also shown to be substantial and significantly different from zero until approximately 1.7kb downstream of the $5'$ splice site.

In Chapter 5 the analysis in Chapter 4 was extended. A larger dataset of over 6,000 mouse-rat gene orthologues was collected and the coding and noncoding regions extracted and aligned. In this chapter transposable elements that were inserted in the last common ancestor of mouse and rats were used as a neutral standard. Specifically, any transposable elements identified by the program RepeatMasker in the alignment of the coding and surrounding noncoding regions were used as neutral standards for the gene orthologue in question. In the estimation of selective constraint in noncoding regions, all repetitive DNA was excluded from the analysis. The results of this chapter suggested that selective constraint in the majority of intergenic DNA is low, but extensive. In contrast to the results presented in Chapter 4, selective constraint in intergenic DNA appears to plateau at approximately 8%, 5kb from transcription start and stop points of adjacent genes. Thus, our results suggest that there is low selective constraint in intergenic DNA located considerable distances ($\sim$50kb) from genic regions. Our results also suggest a relationship between intron length and selective constraint in murid introns. This has been previously observed in

Drosophila (Haddrill et al., 2005; Halligan and Keightley, accepted), but this is the first time this effect has been quantified in a mammalian species. Thus it appears that, whilst introns longer than ~6kb appear to be moderately selectively constrained, introns shorter than this are evolving at approximately the same rate as intronic transposable elements. The differences between the results of Chapters 4 and 5 can be explained as follows. Firstly, in Chapter 4 all noncoding sequence, repetitive and nonrepetitive, was analysed, whilst in Chapter 5, repetitive DNA was excluded from the estimation of selective constraint in noncoding DNA. This revealed higher selective constraint in nonrepetitive intergenic DNA than estimated in Chapter 4. In addition, the results presented in Chapter 5 suggest an effect of transcription on the mutation rate. Thus it appears that fastest evolving intronic sites provide an inadequate model of neutral evolution in intergenic DNA and, for this reason, intergenic selective constraint was underestimated in Chapter 5. At a genomic level, it was estimated that murid noncoding DNA contains over three times as many constrained bases as coding sequence.

The results presented in Chapters 4 and 5 have a number of implications. Firstly, the data show substantial, selectively constrained sequence located throughout murid intergenic DNA. It is likely that at least some of these constrained sites are the same as those identified by multiple cross species comparisons in previous studies (e.g. Dermitzakis et al., 2002, 2003; Siepel et al., 2005). These results provide a quantitative estimate of the evolutionary importance of much noncoding DNA. Secondly, the relationship between intron length and selective constraint suggests that functional motifs are preferentially located in longer introns. Why this should be the case remains a mystery. Finally, the results in Chapter 5 suggest a small, but significant effect, of transcription on the local mutation rate, possibly as a result of transcription coupled repair.

Future Work

A primary objective with respect to the results presented in Chapter 2 is the resolution with the findings of Subramanian and Kumar (2003). There are a number of possible reasons for the discrepancy of their results with the predictions of our analysis, the resolutions of which require the identification of the source genes from which their alignments are derived. One possible

extension of the work presented in Chapter 2 would be to develop an analytical model which explicitly describes the evolution of CpG frequencies with time. Although simulations were used in this chapter to address this question, an analytical solution is more attractive as the influence of a variety of parameters can be easily examined. In addition, exact, rather than approximate, CpG equilibrium frequencies can be obtained. It remains to be seen whether any of the models proposed to estimate CpG and non CpG nucleotide substitution rates (Arndt et al., 2003a; Lunter and Hein, 2004) can be adapted for this purpose.

The results of Chapter 3 suggest substantial intra-chromosomal mutational variation. A potential issue is that rates of nucleotide substitution were used to make inferences about the mutation rate. This is problematic if significant numbers of transposable elements are evolving non-neutrally. One way to avoid this would be to test whether patterns of polymorphism support the conclusion of a 1Mb mutational scale and show larger intra- rather than inter-chromosomal variation. This would be possible in humans, where well annotated, genome-wide polymorphism data is publicly available. In addition to this, it is clear that the estimate of mutational scale in murids is a genome-wide average and may vary considerably. One possible extension of this work could therefore be to partition this variation further, into biologically meaningful locations. For example it would be interesting to investigate whether much mutational variation can be explained by differences between centromeric and telomeric regions. This analysis could perhaps also extend to "synteny blocks" (i.e. those regions in which gene order is conserved between mouse and rat) as these have been previously suggested as the biological "unit" of mutational variation (Webster et al., 2004). A further extension would be to investigate whether any relationship exists between replication origins or chromatin domains and nucleotide substitution rate. Another potential extension would be to examine the relationship between various other genomic variables, such as gene density, with mutational variation. This could possibly reveal whether different mutational forces operate in those regions that are gene dense versus those that gene poor. In part prompted by the results of Chapter 5, it would also be interesting to determine whether much mutational variation could be explained by transcription coupled repair. This could be addressed by investigating whether any difference exists between those transposable elements that are located within genes expressed in the germline, and those that are located either in intergenic DNA or in genes not expressed in the germline. As demonstrated in Chapter 5, this is potentially a significant effect

and could explain substantial mutational variation. Finally, again with reference to the results presented in Chapter 5, it would be interesting to investigate the level of CpG hypermutability upon the X chromosome. Collectively the results presented in this thesis suggest that CpG dinucleotides may be less mutable on the X chromosome than on the autosome. This may results from differences in the methylation level of the X chromosome, and perhaps represent a selective lowering of the mutation rate on the X. This could be done by estimating the level of methylation of CpGs in repetitive DNA and short introns, as both are likely to be evolving close to neutrally, but the methylation frequency may differ between repetitive and non repetitive regions (Meunier et al., 2005).

The results of Chapters 4 and 5 raise a number of interesting questions. Chiefly among them is the purpose and function of the selectively constrained noncoding DNA. One broad way to address this question would be to investigate the relationship (or lack thereof) between selective constraint of noncoding regions the level of gene expression. A number of previous results suggest that a relationship is likely to exist. First, the findings of Castillo-Davis et al. (2002) suggest a that intron length is negatively correlated with gene expression level. Thus it may be that highly expressed genes have relatively few noncoding motifs that regulate their expression level. Second, the results of Nelson et al. (2004) suggest that, in *Drosophila* and *Caenorhabditis*, regulatory complexity is related to noncoding DNA length. Given that there is clearly some relationship between the length and the level of selective constraint in noncoding DNA (Haddrill et al., 2005; Halligan and Keightley, accepted), it would be interesting to investigate whether a relationship between gene expression level or complexity and selective constraint exists within mammals.

The assumption of neutrality of the majority of transposable element DNA in murids is crucial to many aspects of this thesis. It is important that this assumption be tested. In particular, even if most repetitive DNA is evolving neutrally, biased gene conversion may also play an important role in governing the evolutionary rate in transposable elements (Webster et al., 2005). One way to address this question would be to compare fixation probabilities of new mutations within repetitive DNA, with adjacent, nonrepetitive DNA. These could be estimated using combined polymorphism and divergence data from a well annotated species pair, potentially human and chimpanzee. One approach would be to test whether certain types of substitution (e.g. AT→GC)

become fixed more often than expected from the numbers of the corresponding polymorphism. Significant departures from that expected from the pattern of polymorphism could indicate non neutral evolution. A further question to address would be how patterns of polymorphism and divergence differ between repetitive and nonrepetitive DNA. By using non repetitive DNA from short introns, and fragmentary, recently-inserted transposable elements the effects of natural selection could potentially be minimised.

A further caveat to much of the work in this thesis is that the assumption of negligible rates of adaptive evolution prevails throughout. Although data upon genomic rates of adaptive evolution in noncoding DNA are sparse in mammals, recent studies have suggested that adaptive substitutions in noncoding DNA are surprisingly common in *Drosophila* (Andolfatto, 2005). A key extension of the work presented here would be to address this possibility in a mammalian taxon.

Conclusions

The results presented in this thesis relate to the evolutionary importance of noncoding DNA in murids and, to a lesser extent, in mammals as a whole. One primary conclusion of this work is that the functional and evolutionary importance of noncoding DNA has been somewhat underestimated. This conclusion is supported by the results of multiple recent studies revealing extensive between-species conservation of noncoding sequence. One of the major putative roles for functional noncoding DNA is the regulation of gene expression. It has long been suggested that many of the primary phenotypic differences between "higher" organisms may result from differences in the regulation and timing of gene expression, rather than the possession of entirely different complements of proteins (Ohno, 1971; King and Wilson, 1975). If many of the differences between mammalian species are dictated by differences in timing and level of expression, it is not unreasonable to suppose that at least some, and perhaps substantial quantities of, noncoding DNA are adaptively evolving. With the availability of cheap and effective methods of quantifying gene expression in addition to the rapidly growing number of publicly available whole genome sequences, answering this question is now a possibility.

Bibliography

Akashi, H. 1997. Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene* **205**: 269–278.

Akashi, H. and Schaeffer, S. W. 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. *Genetics* **146**: 295–307.

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

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**: 1149–1152.

Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R., and Brenner, S. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. Natl. Acad. Sci. USA* **92**: 1684–1688.

Arndt, P. F., Burge, C. B., and Hwa, T. 2003a. DNA sequence evolution with neighbor-dependent mutation. *J. Comput. Biol.* **10**: 313–322.

Arndt, P. F. and Hwa, T. 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* **21**: 2322–2328.

Arndt, P. F., Hwa, T., and Petrov, D. A. 2005. Substantial regional variation in substitution rates in the human genome: Importance of GC content, gene density, and telomere-specific effects. *J. Mol. Evol.* **60**: 748–763.

Arndt, P. F., Petrov, D. A., and Hwa, T. 2003b. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol.* **20**: 1887–1896.

Bailey, J. A., Carrel, L., Chakravarti, A., and Eichler, E. E. 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc. Natl. Acad. Sci. USA* **97**: 6634–6639.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–1325.

Bell, A. C., West, A. G., and Felsenfeld, G. 2001. Gene regulation - insulators and boundaries: Versatile regulatory elements in the eukaryotic genome. *Science* **291**: 447–450.

Belle, E. M. S., Duret, L., Galtier, N., and Eyre-Walker, A. 2004. The decline of isochores in mammals: An assessment of the gc content variation along the mammalian phylogeny. *J. Mol. Evol.* **58**: 653–660.

Belshaw, R. and Bensasson, D. 2006. The rise and fall of introns. *Heredity* **96**: 201–213.

Bensasson, D., Petrov, D. A., Zhang, D. X., Hartl, D. L., and Hewitt, G. M. 2001. Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Mol. Biol. Evol.* **18**: 246–253.

Bergman, C. M. and Kreitman, M. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11**: 1335–1345.

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**: 1499–1504.

Bird, A. P. 1986. CpG-rich islands and the function of DNA methylation. *Nature* **321**: 209–213.

Blake, C. C. F. 1978. Do genes-in-pieces imply proteins-in-pieces? *Nature* **273**: 267–267.

Blake, R. D., Hess, S. T., and Nicholsontuell, J. 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J. Mol. Evol.* **34**: 189–200.

Boulikas, T. 1993. Nature of DNA sequences at the attachment regions of genes to the nuclear matrix. *J. Cell. Biochem.* **52**: 14–22.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. 1994. *Time Series Analysis: Forecasting and Control. 3rd ed.* Prentice-Hall, New Jersey.

Brack, C. and Tonegawa, S. 1977. Variable and constant parts of immunoglobulin light chain gene of a mouse myeloma cell are 1250 non-translated bases apart. *Proc. Natl. Acad. Sci. USA* **74**: 5652–5656.

Bray, N., Dubchak, I., and Pachter, L. 2003. AVID: A global alignment program. *Genome Res.* **13**: 97–102.

Britten, R. J. 1997. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205**: 177–182.

Britten, R. J. and Kohne, D. E. 1968. Repeated sequences in DNA. *Science* **161**: 529–&.

Brosius, J. 2005. Waste not, want not - transcript excess in multicellular eukaryotes. *Trends Genet.* **21**: 287–288.

Bustamante, C. D., Nielsen, R., and Hartl, D. L. 2002. A maximum likelihood method for analyzing pseudogene evolution: Implications for silent site evolution in humans and rodents. *Mol. Biol. Evol.* **19**: 110–117.

Calabrese, P. and Durrett, R. 2003. Dinucleotide repeats in the *Drosophila* and human genomes have complex, length-dependent mutation processes. *Mol. Biol. Evol.* **20**: 715–725.

Carvalho, A. B. and Clark, A. G. 1999. Genetic recombination - intron size and natural selection. *Nature* **401**: 344–344.

Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V., and Kondrashov, F. A. 2002. Selection for short introns in highly expressed genes. *Nature Genet.* **31**: 415–418.

Cavalier-Smith, T. 1991. Intron phylogeny - A New Hypothesis. *Trends Genet.* **7**: 145–148.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell* **116**: 499–509.

Chamary, J. V. and Hurst, L. D. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: Evidence for selectively driven codon usage. *Mol. Biol. Evol.* **21**: 1014–1023.

Chamary, J. V. and Hurst, L. D. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* **6**.

Chang, B. H. J., Shimmin, L. C., Shyue, S. K., Hewettemmett, D., and Li, W. H. 1994. Weak male-driven molecular evolution in rodents. *Proc. Natl. Acad. Sci. USA* **91**: 827–831.

Chen, F. C., Vallender, E. J., Wang, H., Tzeng, C. S., and Li, W. H. 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.* **92**: 481–489.

Chin, C., Chuang, J. H., and Li, H. 2005. Genome-wide regulatory complexity in yeast promoters: Separation of functionally conserved and neutral sequence. *Genome Res.* **15**: 205–213.

Chuang, J. H. and Li, H. 2004. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol.* **2**: 253–263.

Clark, A. G. 2001. The search for meaning in noncoding DNA. *Genome Res.* **11**: 1319–1320.

Comeron, J. M. 2006. Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. *Proc. Natl. Acad. Sci. USA* **103**: 6940–6945.

Comeron, J. M. and Kreitman, M. 2000. The correlation between intron length and recombination in *Drosophila*: Dynamic equilibrium between mutational and selective forces. *Genetics* **156**: 1175–1190.

Comeron, J. M. and Kreitman, M. 2002. Population, evolutionary and genomic consequences of interference selection. *Genetics* **161**: 389–410.

Commoner, B. 1964. Roles of deoxyribonucleic acid in inheritance. *Nature* **202**: 960–968.

Cowan, N. J., Dobner, P. R., Fuchs, E. V., and Cleveland, D. W. 1983. Expression of human alpha-tubulin genes - interspecies conservation of 3′ untranslated regions. *Mol. Cell. Biol.* **3**: 1738–1745.

Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A., and Stephens, M. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genet.* **36**: 700–706.

De Souza, S. J., Long, M., Kleln, R. J., Roy, S., Lin, S., and Gilbert, W. 1998. Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. USA* **95**: 5094–5099.

Deininger, P. L. and Batzer, M. A. 1999. Alu repeats and human disease. *Mol. Genet. Metab.* **67**: 183–193.

Dermitzakis, E. T., Kirkness, E., Schwarz, S., Birney, E., Reymond, A., and Antonarakis, S. E. 2004. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14**: 852–859.

Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* **6**: 151–157.

Dermitzakis, E. T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B. J., Flegel, V., Bucher, P., Jongeneel, C. V., et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**: 578–582.

Dermitzakis, E. T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C., and Antonarakis, S. E. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences. *Science* **302**: 1033–1035.

Doolittle, W. F. 1978. Genes in pieces: Were they ever together ? *Nature* **272**: 581–582.

Doolittle, W. F. and Sapienza, C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601–603.

Duret, L., Dorkeld, F., and Gautier, C. 1993. Strong conservation of noncoding sequences during vertebrates evolution - potential involvement in posttranscriptional regulation of gene-expression. *Nucleic Acids Res.* **21**: 2315–2322.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162**: 1837–1847.

Ebersberger, I., Metzler, D., Schwarz, C., and Paabo, S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70**: 1490–1497.

Eddy, S. R. 2002. Computational genomics of noncoding RNA genes. *Cell* **109**: 137–140.

Ellegren, H., Smith, N. G. C., and Webster, M. T. 2003. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* **13**: 562–568.

Emorine, L., Kuehl, M., Weir, L., Leder, P., and Max, E. E. 1983. A conserved sequence in the immunoglobulin jk-ck intron - possible enhancer element. *Nature* **304**: 447–449.

Engels, W. R. and Preston, C. R. 1980. Components of hybrid dysgenesis in a wild population of *Drosophila melanogaster*. *Genetics* **95**: 111–128.

Epstein, W. and Beckwith, J. R. 1968. Regulation of gene expression. *Ann. Rev. Biochem.* **37**: 411–436.

Ewens, W. J. 2004. *Mathematical Population Genetics*, volume I. Springer-Verlag, New York, 2nd edition.

Eyre-Walker, A. 1999. Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* **152**: 675–683.

Eyre-Walker, A. and Keightley, P. D. 1999. High genomic deleterious mutation rates in hominids. *Nature* **397**: 344–347.

Fairbrother, W. G., Holste, D., Burge, C. B., and Sharp, P. A. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLos Biol.* **2**: 1388–1395.

Filatov, D. A. 2004. A gradient of silent substitution rate in the human pseudoautosomal region. *Mol. Biol. Evol.* **21**: 410–417.

Filipski, J. 1988. Why the rate of silent codon substitutions in variable within a vertebrate genome. *J. Theor. Biol.* **134**: 159–164.

Frazer, K. A., Sheehan, J. B., Stokowski, R. P., Chen, X. Y., Hosseini, R., Cheng, J. F., Fodor, S. P. A., Cox, D. R., and Patil, N. 2001. Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* **11**: 1651–1659.

Gaffney, D. J. and Keightley, P. D. 2004. Unexpected conserved non-coding DNA blocks in mammals. *Trends Genet.* **20**: 332–337.

Gaffney, D. J. and Keightley, P. D. 2005. The scale of mutational variation in the murid genome. *Genome Res.* **15**: 1086–1094.

Gardiner-Garden, M. and Frommer, M. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**: 261–282.

Gazave, E., Gautier, P., Gilchrist, S., and Bickmore, W. A. 2005. Does radial nuclear organisation influence dna damage? *Chromosome Research* **13**: 377–388.

Ghanem, N., Jarinova, O., Amores, A., Long, Q. M., Hatch, G., Park, B. K., Rubenstein, J. L. R., and Ekker, M. 2003. Regulatory roles of conserved intergenic domains in vertebrate dlx bigene clusters. *Genome Res.* **13**: 533–543.

Gilbert, W. 1978. Why genes in pieces? *Nature* **271**: 501.

Glazko, G. V., Koonin, E. V., Rogozin, I. B., and Shabalina, S. A. 2003. A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.* **19**: 119–124.

Glover, D. M. and Hogness, D. S. 1977. Novel arrangement of 18s and 28s sequences in a repeating unit of *Drosophila melanogaster* rDNA. *Cell* **10**: 167–176.

Gojobori, T., Li, W. H., and Graur, D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**: 360–369.

Gonzalez, C. I., Bhattacharya, A., Wang, W. R., and Peltz, S. W. 2001. Nonsense-mediated mRNA decay in *Saccharomyces cerevisiae*. *Gene* **274**: 15–25.

Gumucio, D. L., Shelton, D. A., Zhu, W., Millinoff, D., Gray, T., Bock, J. H., Slightom, J. L., and Goodman, M. 1996. Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the beta-like globin genes. *Mol. Phylogenet. Evol.* **5**: 18–32.

Haddrill, P. R., Charlesworth, B., Halligan, D. L., and Andolfatto, P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol.* **6**.

Haldane, J. B. S. 1957. The cost of natural selection. *J. Genet.* pp. 511–524.

Halligan, D. L., Eyre-Walker, A., Andolfatto, P., and Keightley, P. D. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res.* **14**: 273–279.

Halligan, D. L. and Keightley, P. D. accepted. Selective constraints in the *Drosophila* genome: Inferences based on a genome-wide interspecies comparison. *Genome Res.* .

Hanawalt, P. C. 1994. Transcription-coupled repair and human-disease. *Science* **266**: 1957–1958.

Hardison, R. C., Roskin, K. M., Yang, S., Diekhans, M., Kent, W. J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**: 13–26.

Hartl, D. L., Moriyama, E. N., and Sawyer, S. A. 1994. Selection intensity for codon bias. *Genetics* **138**: 227–234.

Hasegawa, M., Kishino, H., and Yano, T. A. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial-DNA. *J. Mol. Evol.* **22**: 160–174.

Havilio, M., Levanon, E. Y., Lerman, G., Kupiec, M., and Eisenberg, E. 2005. Evidence for abundant transcription of non-coding regions in the saccharomyces cerevisiae genome. *BMC Genomics* **6**: Art. No. 93.

Hayashida, H. and Miyata, T. 1983. Unusual evolutionary conservation and frequent DNA segment exchange in class-i genes of the major histocompatibility complex. *Proc. Natl. Acad. Sci. USA* **80**: 2671–2675.

Hellmann, I., Zollner, S., Enard, W., Ebersberger, I., Nickel, B., and Paabo, S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13**: 831–837.

Hentze, M. W. and Kulozik, A. E. 1999. A perfect message: RNA surveillance and nonsense-mediated decay. *Cell* **96**: 307–310.

Hess, S. T., Blake, J. D., and Blake, R. D. 1994. Wide variations in neighbor-dependent substitution rates. *J. Mol. Biol.* **236**: 1022–1033.

Hickey, D. A. 1982. Selfish DNA - a sexually-transmitted nuclear parasite. *Genetics* **101**: 519–531.

Hirsch, H. J., Starling, P., and Brachet, P. 1972. Two kinds of insertions in bacterial genes. *Mol. Gen. Genet.* **119**: 191–206.

Huang, S.-W., Friedman, R., Yu, N., Yu, A., and Li, W.-H. 2005. How strong is the mutagenicity of recombination in mammals? *Mol. Biol. Evol.* **22**: 426–431.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., et al. 2005. Ensembl 2005. *Nucleic Acids Res.* **33**: D447–D453.

Hudson, T. J., Church, D. M., Greenaway, S., Nguyen, H., Cook, A., Steen, R. G., Van Etten, W. J., Castle, A. B., Strivens, M. A., Trickett, P., et al. 2001. A radiation hybrid map of mouse genes. *Nature Genet.* **29**: 201–205.

Hwang, D. G. and Green, P. 2004. Bayesian Markov Chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. USA* **101**: 13994–14001.

ICGSC 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

IHGSC 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

IMGSC 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

IRGSC 2004. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature* **428**: 493–521.

Jacobs, L. L. and Pilbeam, D. 1980. Of mice and men - fossil-based divergence dates and molecular clocks. *J. Hum. Evol.* **9**: 551–555.

Jaeger, J. J., Tong, H., and Denys, C. 1986. The age of the *Mus-Rattus* divergence - paleontological data compared with the molecular clock. *C R Acad Sci Ser Ii* **302**: 917–&.

Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9**: 815–824.

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y. T., Roskin, K. M., Chen, C. F., Thomas, M. A., Haussler, D., and Jacob, H. J. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14**: 528–538.

Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. In H. N. Munro, ed., *Mammalian Protein Metabolism*, volume III, pp. 21–132. New York: Academic Press.

Kamal, M., Xie, X., and Lander, E. S. 2006. A large family of ancient repeat elements in the human genome is under strong selection. *Proc. Natl. Acad. Sci. USA* **103**: 2740–2745.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. A., and Gingeras, T. R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**: 916–919.

Keightley, P. and Johnson, T. 2004. MCALIGN: Stochastic alignment of noncoding DNA sequences based on an evolutionary model of sequence evolution. *Genome Res.* **14**: 442–450.

Keightley, P. D. and Eyre-Walker, A. 2000. Deleterious mutations and the evolution of sex. *Science* **290**: 331–333.

Keightley, P. D. and Gaffney, D. J. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl. Acad. Sci. USA* **100**: 13402–13406.

Keightley, P. D., Kryukov, G. V., Sunyaev, S., Halligan, D. L., and Gaffney, D. J. 2005a. Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res.* **15**: 1371–1378.

Keightley, P. D., Lercher, M. J., and Eyre-Walker, A. 2005b. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3**: e42.

Khil, P. P., Oliver, B., and Camerini-Otero, R. D. 2005. X for intersection: retrotransposition both on and off the x chromosome is more frequent. *Trends Genet* **21**: 3–7.

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

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *J. Mol. Evol.* **16**: 111–120.

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

Kimura, M. and Ohta, T. 1971. *Theoretical Aspects of Population Genetics*. Princeton University Press.

King, M.-C. and Wilson, A. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116.

Kirkness, E. F., Bafna, V., Halpern, A. L., Levy, S., Remington, K., Rusch, D. B., Delcher, A. L., Pop, M., Wang, W., Fraser, C. M., et al. 2003. The dog genome: Survey sequencing and comparative analysis. *Science* **301**: 1898–1903.

Kondrashov, A. S. 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**: 435–440.

Kondrashov, A. S. and Crow, J. F. 1993. A molecular approach to estimating the human deleterious mutation-rate. *Hum. Mutat.* **2**: 229–234.

Kondrashov, F. A., Ogurtsov, A. Y., and Kondrashov, A. S. submitted. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *Theor. Pop. Biol.* .

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nature Genet.* **31**: 241–247.

Koop, B. F. and Hood, L. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nature Genet.* **7**: 48–53.

Kumar, S. and Gadagkar, S. R. 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* **158**: 1321–1327.

Kumar, S. and Hedges, S. B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392**: 917–920.

Kumar, S. and Subramanian, S. 2002. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. USA* **99**: 803–808.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lehrman, M. A., Goldstein, J. L., Russell, D. W., and Brown, M. S. 1987. Duplication of 7 exons in ldl receptor gene caused by alu-alu recombination in a subject with familial hypercholesterolemia. *Cell* **48**: 827–835.

Lercher, M. J., Chamary, J. V., and Hurst, L. D. 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res.* **14**: 1002–1013.

Lercher, M. J., Williams, E. J. B., and Hurst, L. D. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18**: 2032–2039.

Lewin, B. 1975a. Units of transcription and translation - relationship between heterogeneous nuclear-RNA and messenger-RNA. *Cell* **4**: 11–20.

Lewin, B. 1975b. Units of transcription and translation - sequence components of heterogeneous nuclear-RNA and messenger-RNA. *Cell* **4**: 77–93.

Lewontin, R. C. and Hubby, J. L. 1966. A molecular approach to study of genic heterozygosity in natural populations .2. Amount of variation and degree of heterozygosity in natural populations of drosophila pseudoobscura. *Genetics* **54**: 595–609.

Li, W.-H. 1997. *Molecular Evolution.* Sinauer Associates, USA.

Li, W. H., Wu, C. I., and Luo, C. C. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**: 58–71.

Liu, M. Y. and Grigoriev, A. 2004. Protein domains correlate strongly with exons in multiple eukaryotic genomes - evidence of exon shuffling? *Trends Genet.* **20**: 399–403.

Logsdon, J. M. and Palmer, J. D. 1994. Origin of introns - early or late? *Nature* **369**: 526–526.

Loots, G. G., Locksley, R. M., Blakespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**: 136–140.

Lu, J. and Wu, C. I. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc. Natl. Acad. Sci. USA* **102**: 4063–4067.

Lunter, G. A. and Hein, J. 2004. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* **20 supp. 1**: i216–i223.

Luo, M. J. and Reed, R. 1999. Splicing is required for rapid and efficient mrna export in metazoans. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **96**: 14937–14942.

Lynch, M. 2002. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA* **99**: 6118–6123.

Lynch, M. and Conery, J. S. 2003. The origins of genome complexity. *Science* **302**: 1401–1404.

Lynch, M. and Kewalramani, A. 2003. Messenger RNA surveillance and the evolutionary proliferation of introns. *Mol. Biol. Evol.* **20**: 563–571.

Lynch, M. and Richardson, A. O. 2002. The evolution of spliceosomal introns. *Curr. Opin. Genet. Devel.* **12**: 701–710.

Majewski, J. and Ott, J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**: 1827–1836.

Makova, K. D., S., Y., and Chiaromonte, F. 2004. Insertions and deletions are male biased too: A whole-genome analysis in rodents. *Genome Res.* **14**: 567–573.

Malcom, C. M., Wyckoff, G. J., and Lahn, B. T. 2003. Genic mutation rates in mammals: Local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol. Biol. Evol.* **20**: 1633–1641.

Marais, G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* **19**: 330–338.

Margulies, E. H., Blanchette, M., Haussler, D., and Green, E. D. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**: 2507–2518.

Marshall, H., Studer, M., Popperl, H., Aparicio, S., Kuroiwa, A., Brenner, S., and Krumlauf, R. 1994. A conserved retinoic acid response element required for early expression of the homeobox gene hoxb-1. *Nature* **370**: 567–571.

Matassi, G., Sharp, P. M., and Gautier, C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**: 786–791.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**: 581–584.

McVean, G. T. and Hurst, L. D. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386**: 388–392.

Meunier, J. and Duret, L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21**: 984–990.

Meunier, J., Khelifi, A., Navratil, V., and Duret, L. 2005. Homology-dependent methylation in primate repetitive DNA. *Proc. Natl. Acad. Sci. USA* **102**: 5471–5476.

Mirsky, A. E. and Ris, H. 1951. The deoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physio.* **34**: 451–462.

Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K., and Yasunaga, T. 1987. Male-driven molecular evolution - a model and nucleotide sequence analysis. *Cold Spring Harbor Symp. Quant. Biol.* **52**: 863–867.

Mouchiroud, D., Donofrio, G., Aissani, B., MacAya, G., Gautier, C., and Bernardi, G. 1991. The distribution of genes in the human genome. *Gene* **100**: 181–187.

Nadeau, J. H. and Taylor, B. A. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA* **81**: 814–818.

Nagylaki, T. 1983. Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. USA* **80**: 6278–6281.

Nei, M. 2005. Selectionism and neutralism in molecular evolution. *Molecular Biology And Evolution* **22**: 2318–2342.

Nelson, C. E., Hersh, B. M., and Carroll, S. B. 2004. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* **5**: R25.

Nixon, J. E. J., Wang, A., Morrison, H. G., McArthur, A. G., Sogin, M. L., Loftus, B. J., and Samuelson, J. 2002. A spliceosomal intron in *Giardia lamblia*. *Proc. Natl. Acad. Sci. USA* **99**: 3701–3705.

Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302**: 413–413.

Oeltjen, J. C., Malley, T. M., Muzny, D. M., Miller, W., Gibbs, R. A., and Belmont, J. W. 1997. Large-scale comparative sequence analysis of the human and murine bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* **7**: 315–329.

Ohno, S. 1971. An argument for the genetic simplicity of man and other mammals. *J. Hum. Evol.* **1**: 651–662.

Ohno, S. 1972. So much"junk" DNA in our genome. *Brookhaven Symp Biol* **23**: 366–370.

Ohno, S. 1988. Universal rule for coding sequence construction - TA CG deficiency TG CT excess. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **85**: 9630–9634.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**: 263–286.

Orgel, L. E. and Crick, F. H. C. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**: 604–7.

Parmley, J. L., Chamary, J. V., and Hurst, L. D. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* **23**: 301–309.

Perry, J. and Ashworth, A. 1999. Evolutionary rate of a gene affected by chromosomal position. *Current Biology* **9**: 987–989.

Pinheiro, J. C. and Bates, D. M. 2000. *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag New York.

Qiu, W. G., Schisler, N., and Stoltzfus, A. 2004. The evolutionary gain of spliceosomal introns: Sequence and phase preferences. *Mol. Biol. Evol.* **21**: 1252–1263.

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLos Comp. Biol.* **1**: 166–175.

R Development Core Team 2004. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 3-900051-07-0.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**: 276–277.

Rosenberg, M. S., Subramanian, S., and Kumar, S. 2003. Patterns of transitional mutation biases within and among mammalian genomes. *Mol. Biol. Evol.* **20**: 988–993.

Roy, A. M., Carroll, M. L., Nguyen, S. V., Salem, A. H., Oldridge, M., Wilkie, A. O. M., Batzer, M. A., and Deininger, P. L. 2000. Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10**: 1485–1495.

Roy, S. W., Fedorov, A., and Gilbert, W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. USA* **100**: 7158–7162.

Ruvinsky, A., Eskesen, S. T., Eskesen, F. N., and Hurst, L. D. 2005. Can codon usage bias explain intron phase distributions and exon symmetry? *Journal Of Molecular Evolution* **60**: 99–104.

Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A., and Kondrashov, A. S. 2001. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17**: 373–376.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M. M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**: 1034–1050.

Siepel, A. and Haussler, D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**: 468–488.

Silva, J. C. and Kondrashov, A. S. 2002. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet.* **18**: 544–547.

Singer, M. F. 1982. SINEs and LINEs - highly repeated short and long interspersed sequences in mammalian genomes. *Cell* **28**: 433–434.

Singh, G. B., Kramer, J. A., and Krawetz, S. A. 1997. Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acids Res.* **25**: 1419–1425.

Sironi, M., Menozzi, G., Comi, G. P., Bresolin, N., Cagliani, R., and Pozzoli, U. 2005. Fixation of conserved sequences shapes human intron size and influences transposon-insertion dynamics. *Trends Genet.* **21**: 484–488.

Smith, N. G. C., Webster, M. T., and Ellegren, H. 2002. Deterministic mutation rate variation in the human genome. *Genome Res.* **12**: 1350–1356.

Sorek, R. and Ast, G. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13**: 1631–1637.

Stein, L. D., Bao, Z. R., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N. S., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLos Biol.* **1**: e22.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M. F., Rifkin, S. A., Hua, S. J., Herreman, T., Tongprasit, W., Barbano, P. E., et al. 2004. A gene expression map for the euchromatic genome of drosophila melanogaster. *Science* **306**: 655–660.

Stoltzfus, A. 2004. Molecular evolution: Introns fall into place. *Curr. Biol.* **14**: R351–R352.

Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M., and Doolittle, W. F. 1994. Testing the exon theory of genes - the evidence from protein structure. *Science* **265**: 202–207.

Storz, G. 2002. An expanding universe of noncoding RNAs. *Science* **296**: 1260–1263.

Subramanian, S. and Kumar, S. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13**: 838–844.

Sved, J. and Bird, A. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. USA* **87**: 4692–4696.

Swift, H. 1950. The constancy of deoxyribose nucleic acid in plant nuclei. *Proc. Natl. Acad. Sci. USA* **36**: 643–650.

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L., and Jones, R. T. 1988. Embryonic epsilon-globin and gamma-globin genes of a prosimian primate (galago-crassicaudatus) - nucleotide and amino-acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**: 439–455.

Tajima, F. and Nei, M. 1984. Estimation of evolutionary distance between nucleotide-sequences. *Molecular Biology And Evolution* **1**: 269–285.

Tamura, K. and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**: 512–526.

Taylor, J., Tyekucheva, S., Zody, M., Chiaromonte, F., and Makova, K. D. 2006. Strong and weak male mutation bias at different sites in the primate genomes: Insights from the human-chimpanzee comparison. *Mol. Biol. Evol.* **23**: 565–573.

Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**: 788–793.

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

Urrutia, A. O. and Hurst, L. D. 2003. The signature of selection mediated by expression on human genes. *Genome Res.* **13**: 2260–2264.

Vendrely, R. and Vendrely, C. 1948. La teneur du noyau cellulaire en acide dsoxyribonucleique a travers les organes, les individus et les especes animales: Techniques et premiers resultats. *Experientia* **4**: 434–436.

Vidal, F., Farssac, E., Tusell, J., Puig, L., and Gallardo, D. 2002. First molecular characterization of an unequal homologous Alu-mediated recombination event responsible for hemophilia. *Thromb. And Haemostasis* **88**: 12–16.

Vinogradov, A. E. 1999. Intron-genome size relationship on a large evolutionary scale. *J. Mol. Evol.* **49**: 376–384.

Waring, M. and Britten, R. J. 1966. Nucleotide sequence repetition - a rapidly reassociating fraction of mouse DNA. *Science* **154**: 791–&.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Webster, M., Smith, N., Lercher, M., and Ellegren, H. 2004. Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol. Biol. Evol.* **21**: 1820–1830.

Webster, M. T., Smith, N. G. C., Hultin-Rosenberg, L., Arndt, P. F., and Ellegren, H. 2005. Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Mol. Biol. Evol.* **22**: 1468–1474.

Willie, E. and Majewski, J. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20**: 534–538.

Wilson, A. C., Carlsoon, S. S., and White, T. J. 1977. Biochemical evolution. *Ann. Rev. Biochem.* **46**: 573–639.

Wolf, Y. I., Kondrashov, F. A., and Koonin, E. V. 2001. No footprints of primordial introns in a eukaryotic genome. *Trends Genet.* **17**: 146–146.

Wolfe, K. H. 1991. Mammalian DNA-replication - mutation biases and the mutation-rate. *Journal Of Theoretical Biology* **149**: 441–451.

Wolfe, K. H., Sharp, P. M., and Li, W. H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283–285.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., et al. 2005. Highly conserved noncoding sequences are associated with vertebrate development. *PLos Biol.* **3**: 116–130.

Yaffe, D., Nudel, U., Mayer, Y., and Neuman, S. 1985. Highly conserved sequences in the 3′ untranslated region of messenger-RNAs coding for homologous proteins in distantly related species. *Nucleic Acids Res.* **13**: 3723–3737.

Yang, Z. B. 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**: 105–111.

Yoder, J. A., Walsh, C. P., and Bestor, T. H. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**: 335–340.

Yu, J., Yang, Z. Y., Kibukawa, M., Paddock, M., Passey, D. A., and Wong, G. K. S. 2002. Minimal introns are not "junk". *Genome Res.* **12**: 1185–1189.

A. CpG simulation program

The following is the C code of the simulation program used in Chapter 2. This basic program was used, with minor modifications, in all analyses presented in Chapter 2. This program can be compiled on a UNIX/Linux platform using gcc.

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <getopt.h>
#include <math.h>

#define phe 1
#define leu 2
#define ile 3
#define met 4
#define val 5
#define ser 6
#define pro 7
#define thr 8
#define ala 9
#define tyr 10
#define his 11
#define gln 12
#define asn 13
#define lys 14
#define asp 15
#define glu 16
#define cys 17
#define trp 18
#define arg 29
#define gly 20
#define stop 21
#define pi 3.14159265358979

FILE *seedfileptr;

/* Mersenne Twister seeds */
int mt_seed,mt_sseed;

struct {
  int bases;
  char *anc,*seq1,*seq2;
} seqs;

struct {
  int CpG,CpG_prone_nCpG,nCpG_prone,total;
} diffs;

struct {
  int CpGs_created,old_CpGs_destroyed,*created,new_CpGs_destroyed;
```

```
} misassign;

int rounds=10,misclassified,maxseq=50000000,int_start;
char *seqtype = NULL;
double scaling;

/*****************************************************************/
/* Beginning of Mersenne Twister random number generation routines */

/* This code is available from http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html */

/*
   A C-program for MT19937, with initialization improved 2002/1/26.
   Coded by Takuji Nishimura and Makoto Matsumoto.

   Before using, initialize the state by using init_genrand(seed)
   or init_by_array(init_key, key_length).

   Copyright (C) 1997 - 2002, Makoto Matsumoto and Takuji Nishimura,
   All rights reserved.

   Redistribution and use in source and binary forms, with or without
   modification, are permitted provided that the following conditions
   are met:

     1. Redistributions of source code must retain the above copyright
        notice, this list of conditions and the following disclaimer.

     2. Redistributions in binary form must reproduce the above copyright
        notice, this list of conditions and the following disclaimer in the
        documentation and/or other materials provided with the distribution.

     3. The names of its contributors may not be used to endorse or promote
        products derived from this software without specific prior written
        permission.

   THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS
   "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT
   LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR
   A PARTICULAR PURPOSE ARE DISCLAIMED.  IN NO EVENT SHALL THE
   COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT,
   INCIDENTAL, SPECIAL,
   EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO,
   PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR
   PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF
   LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING
   NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS
   SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE. */


#include <stdio.h>

/* Period parameters */
#define N 624
#define M 397
#define MATRIX_A 0x9908b0dfUL   /* constant vector a */
```

```
#define UPPER_MASK 0x80000000UL /* most significant w-r bits */
#define LOWER_MASK 0x7fffffffUL /* least significant r bits */

static unsigned long mt[N]; /* the array for the state vector  */
static int mti=N+1; /* mti==N+1 means mt[N] is not initialized */

/* initializes mt[N] with a seed */
void init_genrand(unsigned long s)
{
    mt[0]= s & 0xffffffffUL;
    for (mti=1; mti<N; mti++) {
        mt[mti] =
    (1812433253UL * (mt[mti-1] ^ (mt[mti-1] >> 30)) + mti);
        /* See Knuth TAOCP Vol2. 3rd Ed. P.106 for multiplier. */
        /* In the previous versions, MSBs of the seed affect   */
        /* only MSBs of the array mt[].                        */
        /* 2002/01/09 modified by Makoto Matsumoto             */
        mt[mti] &= 0xffffffffUL;
        /* for >32 bit machines */
    }
}

/* initialize by an array with array-length */
/* init_key is the array for initializing keys */
/* key_length is its length */
/* slight change for C++, 2004/2/26 */
void init_by_array(unsigned long init_key[], int key_length)
{
    int i, j, k;
    init_genrand(19650218UL);
    i=1; j=0;
    k = (N>key_length ? N : key_length);
    for (; k; k--) {
        mt[i] = (mt[i] ^ ((mt[i-1] ^ (mt[i-1] >> 30)) * 1664525UL))
          + init_key[j] + j; /* non linear */
        mt[i] &= 0xffffffffUL; /* for WORDSIZE > 32 machines */
        i++; j++;
        if (i>=N) { mt[0] = mt[N-1]; i=1; }
        if (j>=key_length) j=0;
    }
    for (k=N-1; k; k--) {
        mt[i] = (mt[i] ^ ((mt[i-1] ^ (mt[i-1] >> 30)) * 1566083941UL))
          - i; /* non linear */
        mt[i] &= 0xffffffffUL; /* for WORDSIZE > 32 machines */
        i++;
        if (i>=N) { mt[0] = mt[N-1]; i=1; }
    }

    mt[0] = 0x80000000UL; /* MSB is 1; assuring non-zero initial array */
}

/* generates a random number on [0,0xffffffff]-interval */
unsigned long genrand_int32(void)
{
    unsigned long y;
    static unsigned long mag01[2]={0x0UL, MATRIX_A};
```

```
    /* mag01[x] = x * MATRIX_A  for x=0,1 */

    if (mti >= N) { /* generate N words at one time */
        int kk;

        if (mti == N+1)   /* if init_genrand() has not been called, */
            init_genrand(5489UL); /* a default initial seed is used */

        for (kk=0;kk<N-M;kk++) {
            y = (mt[kk]&UPPER_MASK)|(mt[kk+1]&LOWER_MASK);
            mt[kk] = mt[kk+M] ^ (y >> 1) ^ mag01[y & 0x1UL];
        }
        for (;kk<N-1;kk++) {
            y = (mt[kk]&UPPER_MASK)|(mt[kk+1]&LOWER_MASK);
            mt[kk] = mt[kk+(M-N)] ^ (y >> 1) ^ mag01[y & 0x1UL];
        }
        y = (mt[N-1]&UPPER_MASK)|(mt[0]&LOWER_MASK);
        mt[N-1] = mt[M-1] ^ (y >> 1) ^ mag01[y & 0x1UL];

        mti = 0;
    }

    y = mt[mti++];

    /* Tempering */
    y ^= (y >> 11);
    y ^= (y << 7) & 0x9d2c5680UL;
    y ^= (y << 15) & 0xefc60000UL;
    y ^= (y >> 18);

    return y;
}

/* generates a random number on [0,0x7fffffff]-interval */
long genrand_int31(void)
{
    return (long)(genrand_int32()>>1);
}

/* generates a random number on [0,1]-real-interval */
double genrand_real1(void)
{
    return genrand_int32()*(1.0/4294967295.0);
    /* divided by 2^32-1 */
}

/* generates a random number on [0,1)-real-interval */
double genrand_real2(void)
{
    return genrand_int32()*(1.0/4294967296.0);
    /* divided by 2^32 */
}

/* generates a random number on (0,1)-real-interval */
double genrand_real3(void)
{
```

```
    return (((double)genrand_int32()) + 0.5)*(1.0/4294967296.0);
    /* divided by 2^32 */
}

/* generates a random number on [0,1) with 53-bit resolution*/
double genrand_res53(void)
{
    unsigned long a=genrand_int32()>>5, b=genrand_int32()>>6;
    return(a*67108864.0+b)*(1.0/9007199254740992.0);
}
/* These real versions are due to Isaku Wada, 2002/01/09 added */

/* End of Mersenne Twister random number generator routines */
/*****************************************************************/

void write_seed_mersenne() {
  seedfileptr = fopen("seedfile", "w");
  mt_sseed = genrand_int32();
  fprintf(stderr,"Seed drawn for next replicate using Mersenne twister %d\n",\
 mt_sseed);
  fprintf(seedfileptr, "%d\n",mt_sseed);
  fclose(seedfileptr);
}

int discrete_mersenne(int n) {
  /* int in the range 0-(n-1) */
  int res;
  res = (int)(genrand_real1()*(double)n);
  return(res);
}

int fourfoldaminoacid(char *codon)
{
   if ((aminoacid(codon)==leu)&&(codon[0]=='c')) return 1;
   if (aminoacid(codon)==val) return 1;
   if ((aminoacid(codon)==ser)&&(codon[0]=='t')) return 1;
   if (aminoacid(codon)==pro) return 1;
   if (aminoacid(codon)==thr) return 1;
   if (aminoacid(codon)==ala) return 1;
   if ((aminoacid(codon)==arg)&&(codon[0]=='c')) return 1;
   if (aminoacid(codon)==gly) return 1;
   return 0;
}

double normal_mersenne(double mu, double sdev)
{
  double u1, u2, r;
  u1= genrand_real1();
  if (u1==0.0) u1 = 0.00001;            /*prevent fatal error*/
  if (u1==1.0) u1 = .999999;
  u2= genrand_real1();
  r = sqrt (-2.0*log(u1)) * cos(2.0*pi*u2);
  return(r*sdev + mu);
}

int genpoisson_mersenne(double xm)
```

```
{
   static double sq, alxm, g, oldm = (-1.0);
   double em, t, y;
   if (xm>=200.0)   /* Use normal generator for very high xm */
   {
      return normal_mersenne(0.0, sqrt(xm)) + xm;
   }
   if (xm!=oldm)
   {
      oldm = xm;
      g = exp(-xm);
   }
   em = -1;
   t = 1.0;
   do
   {
      ++em;
      t *= genrand_real1();
   }
   while (t > g);
   return em;
}

int aminoacid(char *codon)
{
      if (strcmp(codon, "aaa")==0) return(lys);
      if (strcmp(codon, "aat")==0) return(asn);
      if (strcmp(codon, "aac")==0) return(asn);
      if (strcmp(codon, "aag")==0) return(lys);
      if (strcmp(codon, "ata")==0) return(ile);
      if (strcmp(codon, "att")==0) return(ile);
      if (strcmp(codon, "atc")==0) return(ile);
      if (strcmp(codon, "atg")==0) return(met);
      if (strcmp(codon, "aca")==0) return(thr);
      if (strcmp(codon, "act")==0) return(thr);
      if (strcmp(codon, "acc")==0) return(thr);
      if (strcmp(codon, "acg")==0) return(thr);
      if (strcmp(codon, "aga")==0) return(arg);
      if (strcmp(codon, "agt")==0) return(ser);
      if (strcmp(codon, "agc")==0) return(ser);
      if (strcmp(codon, "agg")==0) return(arg);
      if (strcmp(codon, "taa")==0) return(stop);
      if (strcmp(codon, "tat")==0) return(tyr);
      if (strcmp(codon, "tac")==0) return(tyr);
      if (strcmp(codon, "tag")==0) return(stop);
      if (strcmp(codon, "tta")==0) return(leu);
      if (strcmp(codon, "ttt")==0) return(phe);
      if (strcmp(codon, "ttc")==0) return(phe);
      if (strcmp(codon, "ttg")==0) return(leu);
      if (strcmp(codon, "tca")==0) return(ser);
      if (strcmp(codon, "tct")==0) return(ser);
      if (strcmp(codon, "tcc")==0) return(ser);
      if (strcmp(codon, "tcg")==0) return(ser);
      if (strcmp(codon, "tga")==0) return(stop);
      if (strcmp(codon, "tgt")==0) return(cys);
      if (strcmp(codon, "tgc")==0) return(cys);
```

```
      if (strcmp(codon, "tgg")==0) return(trp);
      if (strcmp(codon, "caa")==0) return(gln);
      if (strcmp(codon, "cat")==0) return(his);
      if (strcmp(codon, "cac")==0) return(his);
      if (strcmp(codon, "cag")==0) return(gln);
      if (strcmp(codon, "cta")==0) return(leu);
      if (strcmp(codon, "ctt")==0) return(leu);
      if (strcmp(codon, "ctc")==0) return(leu);
      if (strcmp(codon, "ctg")==0) return(leu);
      if (strcmp(codon, "cca")==0) return(pro);
      if (strcmp(codon, "cct")==0) return(pro);
      if (strcmp(codon, "ccc")==0) return(pro);
      if (strcmp(codon, "ccg")==0) return(pro);
      if (strcmp(codon, "cga")==0) return(arg);
      if (strcmp(codon, "cgt")==0) return(arg);
      if (strcmp(codon, "cgc")==0) return(arg);
      if (strcmp(codon, "cgg")==0) return(arg);
      if (strcmp(codon, "gaa")==0) return(glu);
      if (strcmp(codon, "gat")==0) return(asp);
      if (strcmp(codon, "gac")==0) return(asp);
      if (strcmp(codon, "gag")==0) return(glu);
      if (strcmp(codon, "gta")==0) return(val);
      if (strcmp(codon, "gtt")==0) return(val);
      if (strcmp(codon, "gtc")==0) return(val);
      if (strcmp(codon, "gtg")==0) return(val);
      if (strcmp(codon, "gca")==0) return(ala);
      if (strcmp(codon, "gct")==0) return(ala);
      if (strcmp(codon, "gcc")==0) return(ala);
      if (strcmp(codon, "gcg")==0) return(ala);
      if (strcmp(codon, "gga")==0) return(gly);
      if (strcmp(codon, "ggt")==0) return(gly);
      if (strcmp(codon, "ggc")==0) return(gly);
      if (strcmp(codon, "ggg")==0) return(gly);
      fprintf(stderr,"Function aminoacid: Unknown codon %s\n",codon);
      exit(1);
}

int get_seed_mersenne()
{
  /* gets the seed and intialises the generator */
   seedfileptr = fopen("seedfile", "r");
   if (seedfileptr==0) {
      printf("No seedfile, enter seed please ");
      scanf("%ld", &mt_seed);
   } else {
     fscanf(seedfileptr, "%ld", &mt_seed);
     fclose(seedfileptr);
   }
   fprintf(stderr,"Seed read for Mersenne twister %ld\n",mt_seed);
   init_genrand(mt_seed);
}

FILE *openforread(char *str)
{
   FILE *f;
   f = fopen(str, "r");
```

```
    if (f==0)
    {
      fprintf(stderr,"ERROR: File %s not found.\n", str);
      exit(1);
    }
    else fprintf(stderr,"Opened file %s for read.\n", str);
    return(f);
}

FILE *openforwrite(char *str, char *mode)
{
    FILE *f;
    f = fopen(str, mode);
    if (f==0)
    {
      fprintf(stderr,"ERROR: Unable to open file %s for write.\n", str);
       exit(1);
    }
    else fprintf(stderr,"Opened file %s for write, ", str);
    if (mode[0]=='a') fprintf(stderr,"append mode.\n");
    else if (mode[0]=='w') fprintf(stderr,"overwrite mode.\n");
    else {
      fprintf(stderr,"Invalid mode (%c) given in openforwrite\n",mode[0]);
      exit(1);
    }
    return(f);
}

print_seq_fasta(FILE *f, char name[], char seq[], int size) {
  int i,line_count,line_size;
  line_size = 60; //number of bases per line
  line_count = 1;
  if(name != NULL) {
    fprintf(f,">%s\n",name);
  }
  for(i=0;i<size;i++,line_count++) {
    fprintf(f,"%c",seq[i]);
    if(line_count == (line_size)) {
      line_count = 0;
      fprintf(f,"\n");
    }
  }
  fprintf(f,"\n");
}

int generate_fourfold_codon (char *codon, double GC) {
  int n,i;
  char c;
  double u;
  n = discrete_mersenne(31);
  if(n == 0) {
    strcpy(codon,"cta");
  } else if(n == 1) {
    strcpy(codon,"ctt");
  } else if(n == 2) {
    strcpy(codon,"ctc");
```

```
} else if(n == 3) {
  strcpy(codon,"ctg");
} else if(n == 4) {
  strcpy(codon,"gta");
} else if(n == 5) {
  strcpy(codon,"gtt");
} else if(n == 6) {
  strcpy(codon,"gtc");
} else if(n == 7) {
  strcpy(codon,"gtg");
} else if(n == 8) {
  strcpy(codon,"tca");
} else if(n == 9) {
  strcpy(codon,"tct");
} else if(n == 10) {
  strcpy(codon,"tcc");
} else if(n == 11) {
  strcpy(codon,"tcg");
} else if(n == 12) {
  strcpy(codon,"cca");
} else if(n == 13) {
  strcpy(codon,"cct");
} else if(n == 14) {
  strcpy(codon,"ccc");
} else if(n == 15) {
  strcpy(codon,"ccg");
} else if(n == 16) {
  strcpy(codon,"aca");
} else if(n == 17) {
  strcpy(codon,"act");
} else if(n == 18) {
  strcpy(codon,"acc");
} else if(n == 19) {
  strcpy(codon,"acg");
} else if(n == 20) {
  strcpy(codon,"gca");
} else if(n == 21) {
  strcpy(codon,"gct");
} else if(n == 22) {
  strcpy(codon,"gcc");
} else if(n == 23) {
  strcpy(codon,"gcg");
} else if(n == 24) {
  strcpy(codon,"cga");
} else if(n == 25) {
  strcpy(codon,"cgt");
} else if(n == 26) {
  strcpy(codon,"cgc");
} else if(n == 27) {
  strcpy(codon,"cgg");
} else if(n == 28) {
  strcpy(codon,"gga");
} else if(n == 29) {
  strcpy(codon,"ggt");
} else if(n == 30) {
  strcpy(codon,"ggc");
```

```
  } else if(n == 31) {
    strcpy(codon,"ggg");
  } else {
    return 1;
  }
}

int measure_CpG_classes(char *seq,int length, double *CG, \
double *nCG, double *CnG, double *nCnG) {
  int i,j,start,increment;
  char codon[4];
  *CG = *nCG = *CnG = *nCnG = 0.0;
  start = 0;
  increment = 1;
  if(!strcmp(seqtype,"coding")) {
    start = 2;
    increment = 3;
  }
  for(i=start;i<length;i+=increment) {
    if(strcmp(seqtype,"coding")==0) {
      for(j=0;j<3;j++) {
codon[j] = seq[i-2+j];
      }
      codon[j] = '\0';
      if(!fourfoldaminoacid(codon)) {
continue;
      }
/*       printf("%s\n",codon); */
    }
    if(i < length) {
      if(seq[i] == 'a') {
if(seq[i+1] == 't') {
  (*nCnG)++;
} else if(seq[i+1] == 'g') {
  (*nCG)++;
} else if(seq[i+1] == 'c') {
  (*nCnG)++;
} else if(seq[i+1] == 'a') {
  (*nCnG)++;
}
      } else if(seq[i] == 't') {
if(seq[i+1] == 'a') {
  (*nCnG)++;
} else if(seq[i+1] == 'g') {
  (*nCG)++;
} else if(seq[i+1] == 'c') {
  (*nCnG)++;
} else if(seq[i+1] == 't') {
  (*nCnG)++;
}
      } else if(seq[i] == 'g') {
if(seq[i+1] == 'a') {
  (*nCnG)++;
} else if(seq[i+1] == 't') {
  (*nCnG)++;
} else if(seq[i+1] == 'c') {
```

```
  (*nCnG)++;
} else if(seq[i+1] == 'g') {
  (*nCG)++;
}
      } else if(seq[i] == 'c') {
if(seq[i+1] == 'a') {
  (*CnG)++;
} else if(seq[i+1] == 't') {
  (*CnG)++;
} else if(seq[i+1] == 'g') {
  (*CG)++;
} else if(seq[i+1] == 'c') {
  (*CnG)++;
}
      } else {
printf("Error\n");
      }
      /*       if(i < length) { */
      /*  printf("%d %c%c\n",i,seq[i],seq[i+1]); */
      /*       } */
    }
  }
  if(!strcmp(seqtype,"coding")) {
    *CG /= (double)length/3.0;
    *nCG /= (double)length/3.0;
    *CnG /= (double)length/3.0;
    *nCnG /= (double)length/3.0;
  } else {
    *CG /= (double)length;
    *nCG /= (double)length;
    *CnG /= (double)length;
    *nCnG /= (double)length;
  }
}

int measure_CpG(char *seq,int length, double *CpG, int *prone) {
  int i,j,start,increment;
  char codon[4];
  *CpG = 0.0;
  *prone = 0;
  start = 0;
  increment = 1;
  if(!strcmp(seqtype,"coding")) {
    start = 2;
    increment = 3;
  }
  for(i=start;i<length;i+=increment) {
    if(strcmp(seqtype,"coding")==0) {
      for(j=0;j<3;j++) {
codon[j] = seq[i-2+j];
      }
      codon[j] = '\0';
      if(!fourfoldaminoacid(codon)) {
continue;
      }
/*       printf("%s\n",codon); */
```

```
    }
    if(inCGdinucleotide(seq,i,length)) {
      //      printf("%c\n",seq[i]);
      (*CpG)++;
    }
    if(!preCpostG(seq,seq,i,length)) {
      //      printf("%c nCpGprone site\n",seq[i]);
      (*prone)++;
    } else {
      //      printf("%c CpGprone site\n",seq[i]);
    }
  }
  if(!strcmp(seqtype,"coding")) {
    *CpG /= (double)length/3.0;
  } else {
    *CpG /= (double)length;
  }
}

int generate_ancestral(char *seq, int length, double GC) {
  int i,j;
  char n[5];
  char c,codon[4];
  double u;
  if(!strcmp(seqtype,"coding")) {
    for(i=0;i<length/3;i++) {
      generate_fourfold_codon(codon,GC);
      u = genrand_real1();
      if((u >= 0.0)&&(u < GC/2.0)) codon[2] =  'g';
      else if((u >= GC/2.0)&&(u < GC)) codon[2] =  'c';
      else if((u >= GC)&&(u < GC+((1.0-GC)/2.0))) codon[2] =  'a';
      else if((u >= ((1.0-GC)/2.0))&&(u <= 1.0)) codon[2] =  't';
      else {
fprintf(stderr,"Unknown error in generate_ancestral.\
Random uniform u %lf.\nExiting...\n",u);
exit(1);
      }
      for(j=0;j<3;j++) {
/*  // printf("%d %d %d %lf %lf\n",i,j,i*3,fmod(j,3),(i*3)+fmod(j,3)); */
seq[(i*3)+(int)(fmod(j,3))] = codon[j];
// printf("%c\n",seq[(i*3)+(int)(fmod(j,3))]);
      }
    }
    seq[length] = '\0';
  } else {
    for(i=0;i<length;i++) {
      u = genrand_real1();
      if((u >= 0.0)&&(u < GC/2.0)) c = 'g';
      else if((u >= GC/2.0)&&(u < GC)) c = 'c';
      else if((u >= GC)&&(u < GC+((1.0-GC)/2.0))) c = 'a';
      else if((u >= ((1.0-GC)/2.0))&&(u <= 1.0)) c = 't';
      else {
fprintf(stderr,"Unknown error in generate_ancestral.\
Random uniform u %lf.\nExiting...\n",u);
exit(1);
      }
```

```
      seq[i] = c;
    }
    seq[i] = '\0';
  }
  return 0;
}

int preCpostG(char *seq1, char *seq2, int ind, int bases) {
  int i,gap_flag=0;
  if(ind>0) { //not at beginning
    //    printf("%d %c,%c\n",ind,seq[ind],seq[ind-1]);
    if((seq1[ind-1] != '-')&&(seq2[ind-1] != '-')) {
      if((seq1[ind - 1]=='c')||(seq2[ind - 1]=='c')) {
return 1;
      }
    }
  }
  if(ind<bases-1) { //not at end
    //    printf("%d %c,%c\n",ind,seq[ind],seq[ind+1]);
    if((seq1[ind+1] != '-')&&(seq2[ind+1] != '-')) {
      if((seq1[ind + 1]=='g')||(seq2[ind + 1]=='g')) {
return 1;
      }
    }
  }
  return 0;
}

int inCGdinucleotide(char *seq, int ind, int bases) {
   int i,res=0;
   char b;
   b = seq[ind];
   if (b == 'c') {
     if(ind<bases) { //not at end
       if (seq[ind+1]=='g') {
 return 1;
       }
     }
   } else if (b == 'g') {
     if(ind>0) {
       if (seq[ind-1]=='c') {
 return 1;
       }
     }
   }
   return 0;
}

int change_base(char *seq, int loc, \
char *new_base, int CpG_status, int bases, int *i) {
  char old_base;
  double rn,trans;
  old_base = seq[loc];
  rn = genrand_real1();
  if(CpG_status) {
    trans = 0.333333333333333333;
```

```
  } else {
    trans = 0.333333333333333333;
  }
  if(old_base == 'a') {
    if((rn >= 0)&&(rn < trans)) (*new_base) = 'g';
    if((rn >= trans)&&(rn < 2.0*trans)) (*new_base) = 't';
    if((rn >= 2.0*trans)&&(rn < 1.0)) (*new_base) = 'c';
  }
  if(old_base == 't')  {
    if((rn >= 0)&&(rn < trans)) (*new_base) = 'c';
    if((rn >= trans)&&(rn < 2.0*trans)) (*new_base) = 'g';
    if((rn >= 2.0*trans)&&(rn < 1.0)) (*new_base) = 'a';
  }
  if(old_base == 'g')  {
    if((rn >= 0)&&(rn < trans)) (*new_base) = 'a';
    if((rn >= trans)&&(rn < 2.0*trans)) (*new_base) = 't';
    if((rn >= 2.0*trans)&&(rn < 1.0)) (*new_base) = 'c';
  }
  if(old_base == 'c')  {
    if((rn >= 0)&&(rn < trans)) (*new_base) = 't';
    if((rn >= trans)&&(rn < 2.0*trans)) (*new_base) = 'a';
    if((rn >= 2.0*trans)&&(rn < 1.0)) (*new_base) = 'g';
  }
  seq[loc] = (*new_base);
  if(!strcmp(seqtype,"ncoding")) {
    //    printf("%d %d %c->%c \n",loc,bases,old_base,*new_base);
    if(CpG_status == 1) {
      //      printf("CpG destroyed\n");
      if((misassign.created[loc])&&(misassign.created[loc] == 1)) {
misassign.new_CpGs_destroyed++;
      } else {
misassign.old_CpGs_destroyed++;
      }
    }
    if(inCGdinucleotide(seq,loc,bases)) {
      /*       printf("CpG created\n"); */
      misassign.CpGs_created++;
      misassign.created[loc] = 1;
    }
  }
  //  printf("%d rn %lf\told_base %c\tnew_base %c\n",loc,rn,old_base,*new_base);
}

int mutate(char *seq, int bases, int gens, double C_2_nC, \
   double GC, double CpG, double K) {
  int i,n,m,k,mut_flag=0,flag=0;
  char old_codon[4],new_codon[4],old_base,new_base;
  double exp,obs,u,CG,nCG,CnG,nCnG;
  FILE *comp;
  if(strcmp(seqtype,"coding")==0) {
    exp = (K/(2.0*3))*(double)bases;
  } else {
    exp = (K/2.0)*(double)bases;
  }
  obs = genpoisson_mersenne(exp);
  printf("Expected nmuts %1.2f Observed nmuts %1.3f\n",exp,obs);
```

```
  for(i=0;i<obs;) {
    n = discrete_mersenne(bases);
    mut_flag = 0;
    if(inCGdinucleotide(seq,n,bases)) {
      if(!strcmp(seqtype,"coding")) {
for(m=0,k=n-(int)fmod(n,3);m<3;k++,m++) {
  old_codon[m] = seq[k];
}
old_codon[3] = '\0';
strcpy(new_codon,old_codon);
change_base(new_codon,(int)fmod(n,3),&new_base,1,bases,&i);
// change_base(new_codon,(int)fmod(n,3),&new_base);
if(aminoacid(old_codon) == aminoacid(new_codon)) {
  seq[n] = new_base;
  diffs.CpG++;
  //   printf("%s %s\n",old_codon,new_codon);
}
      } else {
old_base = seq[n];
change_base(seq,n,&new_base,1,bases,&i);
// change_base(seq,n,&new_base);
diffs.CpG++;
      }
      i++;
    } else {
      u = genrand_real1();
      if(u < 1.0/C_2_nC) {
if(!strcmp(seqtype,"coding")) {
  for(m=0,k=n-(int)fmod(n,3);m<3;k++,m++) {
    old_codon[m] = seq[k];
  }
  old_codon[3] = '\0';
  strcpy(new_codon,old_codon);
  change_base(new_codon,(int)fmod(n,3),&new_base,0,bases,&i);
  if((aminoacid(old_codon) == aminoacid(new_codon))&&((int)fmod(n,3)==2)) {
    seq[n] = new_base;
    mut_flag++;
    //   printf("%s %s\n",old_codon,new_codon);
  }
} else {
  old_base = seq[n];
  change_base(seq,n,&new_base,0,bases,&i);
  mut_flag++;
}
i++;
      }
    }
    if(mut_flag) {
      if(preCpostG(seq,seq,n,bases)) {
diffs.CpG_prone_nCpG++;
      } else {
diffs.nCpG_prone++;
      }
    }
  }
}
```

```
int main (int argc, char *argv[]) {
  int i,stat,length=0,nCpG_prone_anc,null_diffs,arrowflag=0,ances_from_file=0;
  char c,name[100];;
  double GC=-1.0,CpG1=0.0,CpG2=0.0,CpG_anc,\
    C_2_nC,CpG_elevation=0,K=-1.0,CG,nCG,CnG,nCnG;
  FILE  *outf,*resf,*nullf,*ancf;
  if(argc-1) {
    while ((c = getopt(argc, argv, "g:l:r:t:e:c:s:k:")) != -1) {
      switch(c) {
      case 'g':
GC = atof(optarg);
break;
      case 'l':
length = atoi(optarg);
break;
      case 't':
seqtype = optarg;
break;
      case 'e':
CpG_elevation = atof(optarg);
break;
      case 's':
scaling = atof(optarg);
break;
      case 'k':
K = atof(optarg);
break;
      }
    }
  }
  if(GC<0) {
    fprintf(stderr,"Ancestral GC content not specified (-g switch).\nExiting...\n");
    exit(1);
  }
  if(length < 1) {
    fprintf(stderr,"Sequence length not specified (-l switch).\nExiting...\n");
    exit(1);
  }
  if(seqtype == NULL) {
    fprintf(stderr,"Sequence type not specified (-t switch).\nExiting...\n");
    exit(1);
  }
  if(!CpG_elevation) {
    fprintf(stderr,"CpG_elevation not specified (-e switch).\nExiting...\n");
    exit(1);
  }
  if((!strcmp(seqtype,"coding"))&&(fmod(length,3))) {
    fprintf(stderr,"Coding sequence length not a multiple of 3\nExiting...\n");
    exit(1);
  }
  if(K<0) {
    fprintf(stderr,"Tree length not specified (-k switch)\nExiting...\n");
    exit(1);
  }
  get_seed_mersenne();
```

```
 nullf = openforwrite("sim.null.fsa","w");
 resf = openforwrite("simulate_CpG.out","w");
 outf = openforwrite("sim.fsa","w");
 ancf = openforwrite("sim.null.fsa","w");
 diffs.CpG = 0;
 diffs.CpG_prone_nCpG = 0;
 diffs.nCpG_prone = 0;
 misassign.CpGs_created = misassign.old_CpGs_destroyed\
= misassign.new_CpGs_destroyed = 0;
 misassign.created = (int *) malloc((length+1)*sizeof(int));
 seqs.anc = (char *) malloc(maxseq*sizeof(char));
 if(ances_from_file==1) {
   if(strcmp(seqtype,"coding")==0) {
     ancf = openforread("CDS.fsa");
     read_seq_fasta(seqs.anc,name,&length,&arrowflag,ancf,maxseq);
     fclose(ancf);
   } else {
     ancf = openforread("INT.fsa");
     read_seq_fasta(seqs.anc,name,&length,&arrowflag,ancf,maxseq);
     fclose(ancf);
   }
 } else {
   generate_ancestral(seqs.anc,length,GC);
 }
 realloc(seqs.anc,length+1*sizeof(char));
 seqs.seq1 = (char *) malloc(length+1*sizeof(char));
 seqs.seq2 = (char *) malloc(length+1*sizeof(char));
 measure_CpG(seqs.anc,length,&CpG_anc,&nCpG_prone_anc);
 /*   Simulate null model with no CpG hypermutability */
 strcpy(seqs.seq1,seqs.anc);
 strcpy(seqs.seq2,seqs.anc);
 C_2_nC = 1.0;
 mutate(seqs.seq1,length,rounds,C_2_nC,GC,CpG_anc,K);
 mutate(seqs.seq2,length,rounds,C_2_nC,GC,CpG_anc,K);
 print_seq_fasta(ancf,"ancestral",seqs.anc,length);
 null_diffs = diffs.CpG_prone_nCpG + diffs.nCpG_prone + diffs.CpG;
 diffs.CpG_prone_nCpG = diffs.nCpG_prone = diffs.CpG = 0;
 print_seq_fasta(nullf,"seq1",seqs.seq1,length);
 print_seq_fasta(nullf,"seq2",seqs.seq2,length);
 /*    Simulate CpG hypermutability      */
 strcpy(seqs.seq1,seqs.anc);
 strcpy(seqs.seq2,seqs.anc);
 C_2_nC = CpG_elevation;
 mutate(seqs.seq1,length,rounds,C_2_nC,GC,CpG_anc,K);
 mutate(seqs.seq2,length,rounds,C_2_nC,GC,CpG_anc,K);
 print_seq_fasta(outf,"seq1",seqs.seq1,length);
 print_seq_fasta(outf,"seq2",seqs.seq2,length);
 diffs.total = diffs.CpG_prone_nCpG + diffs.nCpG_prone + diffs.CpG;
 if(!strcmp(seqtype,"coding")) {
   printf("%1.1f\t%1.4f\t%d\t%d\t%d\t%d\t%lf\t%lf\n",\
  GC,CpG_anc,diffs.total,diffs.nCpG_prone,diffs.CpG_prone_nCpG,\
  diffs.CpG,(double)diffs.total/((double)length/3.0),\
  (double)null_diffs/((double)length/3.0));
   fprintf(resf,"%1.5f\t%1.10f\t%1.10f\n",\
   CpG_anc,(double)diffs.total/((double)length/3.0),\
   (double)null_diffs/((double)length/3.0));
```

```
  } else {
    printf("nCG\tCG\tTotal\n");
    printf("%d %d %d\n",diffs.nCpG_prone+diffs.CpG_prone_nCpG,\
       diffs.CpG,diffs.nCpG_prone+diffs.CpG_prone_nCpG+diffs.CpG);
    fprintf(resf,"%1.1f\t%1.4f\t%d\t%d\t%d\t%d\t%d\t%lf\t%lf\n",\
    GC,CpG_anc,nCpG_prone_anc,diffs.total,diffs.nCpG_prone,\
    diffs.CpG_prone_nCpG,diffs.CpG,(double)diffs.total/(double)length,\
    (double)null_diffs/(double)length);
  }
  fclose(outf);
  fclose(nullf);
  fclose(resf);
  fclose(ancf);
  write_seed_mersenne();
}
```

B. Publications

Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents

Peter D. Keightley* and Daniel J. Gaffney

Ashworth Laboratories, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom

Edited by James F. Crow, University of Wisconsin, Madison, WI, and approved September 12, 2003 (received for review May 29, 2003)

Selection against deleterious mutations imposes a mutation load on populations because individuals die or fail to reproduce. In vertebrates, estimates of genomic rates of deleterious mutations in protein-coding genes imply the existence of a substantial mutation load, but many functionally important regions of the genome are thought to reside in noncoding DNA, and the contribution of noncoding DNA to the mutation load has been unresolved. Here, we infer the frequency of deleterious mutations in noncoding DNA of rodents by comparing rates of substitution at noncoding nucleotides with rates of substitution at the fastest evolving intronic sites of adjacent genes sampled from the whole genome sequences of mouse and rat. We show that the major elements of selectively constrained noncoding DNA are within 2,500 bp upstream and downstream of coding sequences and in first introns. Our estimate of the genomic deleterious point mutation rate for noncoding DNA (0.22 per diploid per generation) is similar to that for coding DNA. Mammalian populations therefore experience a substantial genetic load associated with selection against deleterious mutations in noncoding DNA. Deleterious mutations in noncoding DNA have predominantly quantitative effects and could be an important source of the burden of complex genetic disease variation in human populations.

selective constraints | intron | intergenic DNA

Selection against deleterious mutations leads to a mutation load at the population level, because individuals die prematurely or have reduced fertility (1, 2). The mutation load can be defined as the proportion of individuals that are selectively eliminated (individuals that undergo "genetic death"; ref. 3) and depends critically on the genomic deleterious mutation rate, U. For example, under a multiplicative model the load is $1 - e^{-U}$ (where U is the mutation rate per diploid; ref. 4). The mutation load also depends on the manner in which mutations interact with one another between and within loci (4), and on population structure and system of mating (5), and can be reduced, for example, if mutations interact synergistically (4). The genomewide rate for mutations in coding DNA has been estimated on the basis of the fraction of conserved nucleotides at amino acid sites of protein-coding genes (6–8). There is a strong, positive correlation between generation time of a species and U (8), and U in long-lived taxa such as hominids is likely to exceed one event per generation (6, 7). However, the contribution of mutations in noncoding DNA to the genomewide deleterious mutation rate is an unresolved issue, because it has been difficult to relate function with DNA sequence, and, until recently, relevant data have not been available.

Protein-coding gene sequences comprise only a very small proportion of the total genomic content in mammals, most other vertebrates, many invertebrates, and most plants (9). For example, protein-coding sequences are thought to account for only ≈1.5% of the genomes of both humans and mice (10, 11). As much as 45% of the euchromatic portion of mammalian genomes consists of the remnants of transposable element insertions (10) that are only occasionally coopted for use by the host organism. Much of the remainder of the genome consists of unique intergenic and intronic DNA sequences, and motifs that are critical for regulating gene expression reside in these regions. Quantification of the degree of between-species evolutionary conservation is one way of searching for such regulatory regions (12). Over evolutionary time scales, directional selection is expected to drive the efficiency of a functional stretch of the genome toward an adaptive optimum, and most non-neutral mutations within it are expected to be deleterious. The between-species evolutionary divergence of functionally important regions is therefore expected to be lower than the divergence of neutral segments having similar mutation rates; those mutations in functional regions with selection coefficients higher than the reciprocal of the effective population size almost never become fixed between species (13).

A general approach to identify functionally important regions in the genome and to quantify the fraction of deleterious mutations is to search for segments of the genome having lower between-species levels of divergence than the average for the genome or than a linked putatively neutral sequence (14). Previous attempts to quantify the fraction of conserved nucleotides have relied on searching for blocks of DNA sequences that are conserved between distantly related taxa (15–18). However, there are at least two difficulties with this approach. First, estimation of noncoding DNA sequence alignment by heuristic methods can be biased if the true pattern of insertion/deletion (indel) events is unknown (19), and second, variability across the genome in the mutation rate can generate variation in conservation that is unrelated to functional constraint (12).

Here, we attempt to quantify the functional constraints on noncoding DNA in rodents by using the recently released genome sequences of mouse and rat by comparing rates of substitution in segments of noncoding DNA with rates of substitution at the fastest evolving intronic (FEI) sites of adjacent genes. We determined empirically that the intronic sites showing the fastest rates of evolution are those nucleotides not close to exon boundaries (i.e., not close to intron splice control regions) and outside first introns. We confine our analysis of constraints to those sites that are unlikely to have been ancestrally part of a CG dinucleotide, because such sites are close to saturation between mouse and rat. The whole genome sequence of the mouse has recently been published (11), and the whole genome sequence of the rat is publicly available on GenBank at seven to eight times coverage. Mouse and rat are sufficiently closely related that it is possible to be confident in the orthology of noncoding DNA sequences. We use a probabilistic method for sequence alignment (P.D.K. and T. Johnson, unpublished work), based on an evolutionary model of indel evolution. We focus our analysis on noncoding DNA sequences associated with well-annotated loci and use estimates of levels of constraint to infer the fraction of deleterious mutations in noncoding DNA.

Methods

Compilation of DNA Sequence Data. We compiled coding and adjacent noncoding DNA sequences from orthologous mouse

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: FEI, fastest evolving intronic.

*To whom correspondence should be addressed. E-mail: p.keightley@ed.ac.uk.

www.pnas.org/cgi/doi/10.1073/pnas.2233252100

and rat loci by random sampling from their respective whole genome sequence assemblies in GenBank. Using the mouse genome as the reference, we randomly selected chromosomes in proportion to their lengths in Mb, then randomly selected positions within chromosomes from a uniform distribution. The nearest locus to this position for which annotation evidence included at least one complete mRNA sequence in both species was chosen. The first 200 loci were sampled at random irrespective of their distances to the next coding sequences. To increase the sample size of loci with long intergenic regions, we sampled an additional 100 loci for which the annotation in both mouse and rat indicated that the nearest coding sequence was >6 kb away. The coding sequences were aligned by using CLUSTAL (20), and positions of gaps were examined and adjusted if necessary. Sequences of lengths of up to 6,000 bp from 5′ upstream and 3′ downstream regions of coding sequences were extracted, along with the first and last introns and one other randomly selected intron. Some genes contain extremely long introns, particularly those that are lowly expressed (21), so we initially focused our analysis of constraint on the first and last 1,000 bp, if intron length exceeded 2,000 bp, otherwise we analyzed complete intron sequences. A further data set of up to 12,000 bp from first introns was subsequently extracted. Only clearly orthologous intergenic and intronic sequences were analyzed; we used a moving window of 40 bp to check the degree of sequence divergence in putative alignments. In some cases we observed sharp jumps in the divergence to ≈60% (the divergence expected for alignment of nonorthologous mouse-rat sequences), whereas the typical mouse-rat divergence is ≈15%. We interpreted these as being caused by a long insertions or deletions or sequence assembly errors. Such obviously nonhomologous regions were excluded from the analysis.

Sequence Alignment. Noncoding DNA sequences were aligned by using a Monte Carlo alignment procedure, MCALIGN, which searches for the alignment of highest probability based on a specific evolutionary model of noncoding DNA sequence evolution (P.D.K. and T. Johnson, unpublished work). Briefly, the parameters of the model are θ, the rate of indels relative to the rate of nucleotide substitutions, and a vector w, the frequency distribution of indel lengths. These parameters are estimated empirically from other data (see below). The Monte Carlo procedure carries out an uphill search of the parameter space of plausible alignments by accepting or rejecting proposal alignments depending on their relative probabilities. New proposal alignments are generated by a set of indel shuffling routines.

The parameters for the alignment model (θ and w) were estimated from 27 orthologous intron sequences of the closely related mouse species *Mus domesticus*, *Mus spretus*, and *Mus caroli*, for which nucleotide and indel divergences are sufficiently low as to make alignments by heuristic methods practically unambiguous. In comparisons between *M. domesticus* and *M. spretus* (10 loci), θ was 0.188, and between *M. domesticus* and *M. caroli* (8 loci), θ was 0.125. A weighted average estimate of θ = 0.146 was used to parameterize the alignment model. The empirical distribution of indel lengths is shown in Fig. 1. To parameterize the alignment model a value for $w_1 = 0.565$ (the empirical value) was assumed; for $i > 1$, values of w_i were estimated by minimizing the sum of squares about a smoothing function, $w_i = \beta/\alpha^i$, where β is a normalizing constant and α is the smoothing parameter. The estimated value of α was 1.45. We used intronic data to parameterize the alignment model for intergenic DNA, which is an approximation. However, indels occur at a lower frequency in intergenic than intronic DNA, on average, and using this approximation will give alignments very close, on average, to the true alignments for the degree of sequence divergence between mouse and rat (unpublished data).

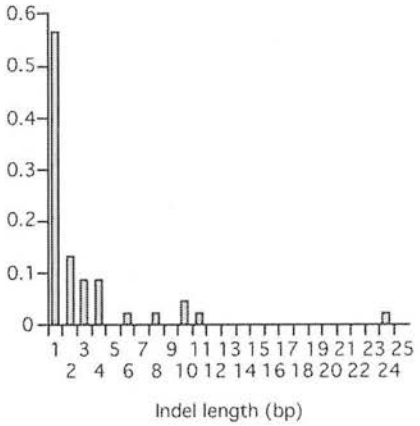

Fig. 1. Frequency distribution of indel length from between-species comparisons of closely related mouse species.

Masking of Microsatellite Repeats. Repeats of type $(XY)_n$, $(XYZ)_n$, $(XYY)_n$, $(XYYY)_n$, and $(XXYYY)_n$, where $n = 5$ in intronic or intergenic regions were excluded from the analysis, because their evolution is not driven by single nucleotide substitutions (22). Sequences adjacent to perfect microsatellite regions showing >80% homology to the specific repeat were also excluded. In addition, it was frequently observed that boundaries of microsatellites contained short stretches of obviously nonhomologous nucleotides, so 5l nucleotides adjacent to microsatellites, where l is the repeat length, were also excluded.

Calculation of Evolutionary Constraint. We calculated constraint in noncoding regions by extending a method previously developed for coding sequences (6). We used substitution rates at FEI sites to predict expected numbers (E) of substitutions in adjacent noncoding sequences (i.e., intergenic DNA, intronic splice sites, or first introns), under the assumption that point mutation rates of each possible kind are equal at FEI sites and the adjacent noncoding DNA sites. We calculated constraint by comparing E to numbers of observed substitutions (O):

$$C = 1 - O/E. \quad [1]$$

There are substantial differences in substitution rates between different kinds of nucleotide (23), so we needed to account for differences in substitution rates between FEI sites and adjacent noncoding regions. For each possible substitution type i = 1.6 (A↔T, A↔C, A↔G, T↔C, T↔G, C↔G), let k_i be the pairwise divergence in the FEI segment, i.e.,

$$k_i = \delta_i/N_i, \quad [2]$$

where d_i is the number of pairwise differences of type i, and N_i is the number of sites at which a change of type i could occur in one step (e.g., for A↔T changes, these sites are A/A, T/T, and T/A). The expected number of substitutions in an adjacent noncoding segment is,

$$E = \sum_{i=1}^{6} k_i M_i, \quad [3]$$

where M_i is the corresponding number of noncoding sites. This model assumes symmetric mutation rates and equivalent base composition in the FEI sites and the noncoding region of interest. However, analysis in which polarity of substitution was assigned via the relative probability of ancestry of each base gave very similar results (data not shown). The method to calculate

Table 1. Proportions of differences at nucleotides within and outside of CG dinucleotides at 4-fold and FEI sites

	Type of nucleotide change					
	A↔T	A↔C	A↔G	T↔C	T↔G	C↔G
Four-fold, within CG	—	0.182 (0.012)	0.468 (0.014)	0.468 (0.015)	0.199 (0.011)	0.149 (0.007)
FEI, within CG	—	0.160 (0.006)	0.385 (0.007)	0.382 (0.007)	0.163 (0.006)	0.105 (0.003)
Four-fold, outwith CG	0.0273 (0.0013)	0.0265 (0.0012)	0.0563 (0.0018)	0.0624 (0.0019)	0.0183 (0.0010)	0.0161 (0.0009)
FEI, outwith CG	0.0293 (0.0005)	0.0315 (0.0006)	0.0866 (0.0010)	0.0823 (0.0010)	0.0321 (0.0006)	0.0276 (0.0006)

Entries are the proportions of nucleotide changes at corresponding categories of sites, where, for example, A↔C sites are A/C, C/A, A/A, or C/C in mouse/rat. Bootstrap SEMs are shown in parentheses.

constraint does not attempt to account for multiple hits. However, simulation results (data not shown) suggest that estimation bias is negligible for nucleotide divergence well in excess of that of mouse-rat (15%) and for substantial differences in GC content. Standard errors and confidence limits for C were calculated by bootstrapping the gene-specific values of O and E 10,000 times.

Calculation of Genomic Deleterious Mutation Rate. The contribution to the deleterious mutation rate from a DNA segment was calculated from the product of average deleterious mutation rate per site (u) and the number of nucleotide sites (s) in the segment. We subdivided sites into two classes: (i) sites preceded by C or followed by G, termed "CG-susceptible," or (ii) all other sites, termed "non-CG-susceptible." A weighted average of contributions from CG-susceptible sites (mutation rate = u_1; number of sites = s_1) and non-CG-susceptible sites (mutation rate = u_2; number of sites = s_2) was taken. The deleterious mutation rate per site (u) was calculated from the product of constraint in the corresponding segment (Eq. **1**) and the nucleotide divergences specific to non-CG-susceptible and CG-susceptible sites, calculated by using Kimura's two-parameter method (9). The contribution of intergenic or intronic DNA to the genomic deleterious mutation rate per diploid (U) was calculated by summing the average contributions of segments:

$$U = Z \sum_{i=1}^{segments} P_i l_i \frac{\sum_{j=1}^{loci} s_{1ij}u_{1ij} + s_{2ij}u_{2ij}}{\sum_{j=1}^{loci} s_{1ij} + S_{2ij}}, \qquad [4]$$

where l_i is the length of a segment (200 bp in the analyses carried out here), P_i is the fraction of loci that actually contain intergenic or intronic DNA in segment i, and Z is a constant to convert between the scale of nucleotide mutation rate and genomic mutation rate per generation: $Z = 35{,}000$ loci $\times$ 0.5 (generations per year)/13 $\times$ 10^6 (approximate age in years of *mus-rattus* divergence; ref. 24). The inclusion of the term P_i was necessary because intergene regions and introns vary in length, and the fraction of loci containing a specific DNA segment declines as the distance in bp from the coding sequence to the start of the segment increases. For intergenic regions, values of P_i were calculated from the first 200 loci sampled (which were assumed to be a random sample with respect to intergene length) and for introns from all loci sampled. Lengths of intergene regions used to calculate P_i were taken as one-half of actual intergene lengths. Mutation rates in CpG islands (regions of the genome, often close to the 5′ end of genes, that are unusually rich in CG dinucleotides) are an order of magnitude lower than most CG sites (25). In the analysis, we assumed that mutation rates at nucleotides within CG-susceptible sites in CpG islands were the same as rates in non-CG-susceptible nucleotides.

Delimiting of CpG Islands. The locations of CpG islands were estimated by using the CpG Plot/CpG Report utilities available from the European Bioinformatics Institute (www.ebi.ac.uk/index.html; ref 26). A CpG island was reported if the observed to expected ratio of C plus G to CpG exceeded 0.6, in regions of GC content >50% (27), in a succession of 10 $\times$ 100-bp windows. Islands of >200 bp only were reported.

Results and Discussion

In a preliminary analysis, we compared nucleotide substitution rates at synonymous sites with rates for various categories of noncoding DNA sites. The analysis showed that introns evolve substantially faster, on average, than intergenic DNA, and that the rate of nucleotide substitution in intron sequences close to exon boundaries is slower than intron sequences in general. The first 1,000 bp of the 5′ region of first introns evolve noticeably slower than introns in general. These preliminary results are not shown, but they are implicit in the results on selective constraints that follow. Based on the preliminary analysis, we used sequences in introns, excluding intron 1, and 6 bp at the 5′ end and 30 bp at the 3′ end of each intron, as the FEI sequences. These sequences are used as standards to infer mutation rates in the subsequent analyses.

Comparison of Proportions of Substitutions in CG Dinucleotide and Non-CG Dinucleotide Sites. Proportions of nucleotide differences at 4-fold degenerate sites and FEI sites are shown in Table 1, split according to whether or not nucleotides are part of CG dinucleotides in either species. The high fraction of differences at CG dinucleotide sites in both 4-fold and noncoding DNA implies that CG dinucleotide sites are close to saturation. Proportions of nucleotide differences within CG dinucleotides are higher at 4-fold sites than FEI sites, whereas proportions outside of CG dinucleotides are higher at FEI sites than 4-folds. However, this

Table 2. Proportions of nucleotide differences at non-CG-susceptible 4-fold and FEI sites

	Type of nucleotide change					
	A↔T	A↔C	A↔G	T↔C	T↔G	C↔G
Four-fold sites	0.0175 (0.0021)	0.0272 (0.0017)	0.0827 (0.0030)	0.0753 (0.0029)	0.0270 (0.0019)	0.0298 (0.0017)
FEI sites	0.0259 (0.0007)	0.0327 (0.0007)	0.0924 (0.0012)	0.0880 (0.0011)	0.0332 (0.0007)	0.0316 (0.0007)

Entries are defined as for Table 1. Bootstrap SEMs are shown in parentheses.

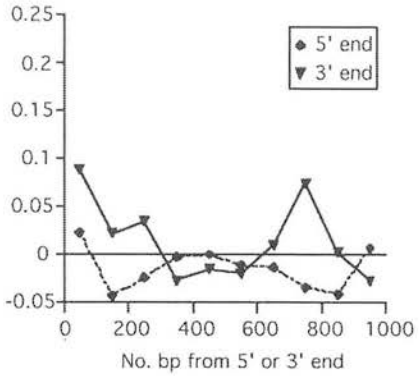

Fig. 2. Evolutionary constraint plotted against distance from the coding sequence (bp) in 100-bp blocks of the 5′ and 3′ ends of FEIs.

pattern is largely caused by ascertainment bias: the high selective constraints at amino acid sites in coding DNA cause ancestral CG sites to be more frequently correctly assigned to the CG category of sites than sites in relatively unconstrained intronic DNA. Conversely, ancestral CG sites in intronic DNA have a high probability of mutation away from CG in both species, so are more often incorrectly assigned to the category of non-CG sites than 4-fold sites. It is therefore inappropriate to simply exclude sites within CG dinucleotides. A less biased procedure was found to be to exclude CG-susceptible sites, those nucleotide sites that are preceded by C or followed by G in either species, and therefore have a high chance of being part of an ancestral CG dinucleotide. Simulations of noncoding DNA evolution including hypermutable CG dinucleotides suggested that such a procedure gives relatively unbiased estimates of constraint for cases of overall nucleotide divergence similar to mouse and rat (results not shown). In all subsequent analyses, this procedure was followed for calculating constraint.

Comparison of Substitution Rates at Non-CG-Susceptible Sites Between 4-Fold Sites and FEI Sites. Outside of CG-susceptible sites, fractions of nucleotide differences at 4-fold sites are consistently lower than at FEI sites (Table 2). It is notable that the fraction of A↔T changes at A/T sites is ≈30% lower at 4-fold sites than at FEI sites. Because A/T sites are four mutational changes from a CG-susceptible site, this finding suggests that the slower rate of substitution at 4-fold sites is unlikely to be a consequence of incorrect assignment of CG dinucleotide status. It is possible that the effect is a consequence of selection, although a role for selection at synonymous sites has been discounted (28). Slower rates of nucleotide substitution at 4-fold sites than noncoding sites have been reported in primates (29, 30).

Evolutionary Constraint in Intronic DNA. In FEIs, the average level of constraint is zero, by definition, because FEIs are assumed in this analysis to be the neutrally evolving standard against which constraint is measured. We tested for variation about this average by calculating mean constraint in 100-bp segments of the FEIs (i.e., the complete FEI data set was used to calculate constraint specific to intronic segments; Fig. 2). Mean constraint is nonsignificantly different from zero along the whole 1,000-bp length at the 5′ end of FEIs and is also nonsignificantly different from zero at the 3′ end of FEIs, with the exception of the first 100 bp at the 3′ end ($P < 0.001$; presumably associated with intronic splice control), and a marginally significantly constrained region at 700–800 bp of the 3′ end ($P = 0.02$). We examined constraint in more detail near 5′ and 3′ splice control regions (Table 3). As expected, there is a strong signal of purifying selection at the dinucleotides adjacent to the 5′ and 3′

Table 3. Estimates of mean selective constraint in intron sequences

Intronic DNA data set	Constraint
5′ bases 1–2	1.0 (0.0)
5′ bases 3–6	0.57 (0.044)
5′ bases 7–10	0.025 (0.076)
3′ bases 1–2	1.0 (0.0)
3′ bases 3–16	0.31 (0.031)
3′ bases 17–30	0.15 (0.040)
Intron 1, 5′ end	0.10 (0.017)
Intron 1, 3′ end	0.0056 (0.016)

In the analysis of intron 1, up to 6,000 bp at the 3′ or 5′ ends were analyzed. If a first intron was <12,000 bp long, the intron sequence was divided equally at the central nucleotide between data sets of 5′ and 3′ sequences. SEMs are shown in parentheses.

splice sites (which are invariant), and in the sequences within 6 bp and ≈30 bp of the 5′ and 3′ ends, respectively, known from previous work to be intimately involved in intron splicing and to be conserved across taxa (31). It has recently been shown that there is higher frequency of transcriptional regulatory sequences in first introns than introns in general (32). Analysis of our data set also supports this observation by revealing constrained sequences in intron 1, located within ≈2 kb of the 5′ end (Table 3 and Fig. 3*A*). The 3′ ends of first introns evolve at a similar rate to FEIs (Table 3).

Evolutionary Constraint in Intergenic DNA. Evolutionary constraint in intergenic regions is moderately strong close to the 5′ and 3′ ends of coding sequences, then drops off surprisingly slowly as a function of distance from the gene (Fig. 3*B*). Some 5′ and 3′ intergenic regions are extremely strongly conserved: ≈5% of loci contain runs of 100 bp within 200 bp of the start or stop codon that are identical between mouse and rat (average sequence

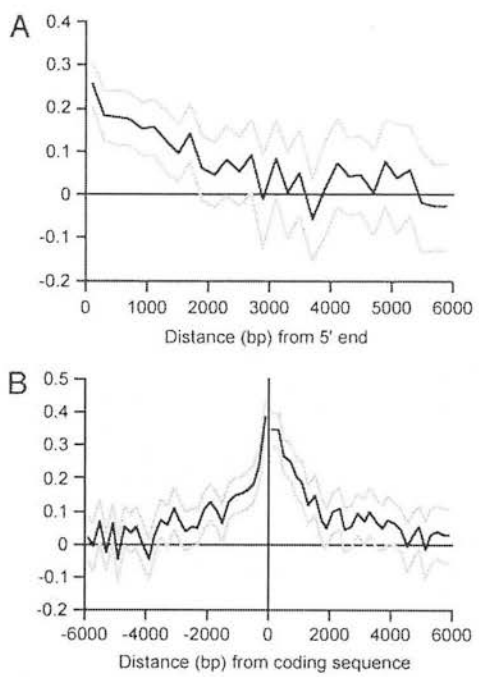

Fig. 3. Evolutionary constraint plotted against distance from the coding sequence (bp) calculated in 200-bp blocks of the 5′ end of first introns (*A*) and in intergenic regions (*B*). The upper and lower 95% confidence limits are shown in light gray.

Table 4. Estimates of selective constraint in coding, intronic, and intergenic DNA of rodents, and contributions to the genomic deleterious mutation rate (U) per diploid genome per generation

DNA category	Nucleotide sites per locus	Mean constraint, SEM	Contribution to U
Coding	1,125*	0.87 (0.009)	0.22¶
5′ intronic splice regions	44.4†	0.73 (0.027)	0.0071
3′ intronic splice regions	222†	0.29 (0.024)	0.012
Intron 1, 5′ end	3,307‡	0.10 (0.017)	0.049
5′ intergenic	5,596§	0.093 (0.013)	0.074
3′ intergenic	5,271§	0.12 (0.015)	0.079

Estimates were made under the assumption that there are 35,000 protein coding loci in the mouse genome (10, 11).

*The average length of rodent coding sequences is ≈1,500 nt, and about three-quarters of sites in coding sequences generate an amino acid change if mutated.

†Blocks totaling 6 and 30 nt near 5′ and 3′ splice junction sites, respectively, show significant evidence of selective constraint (Table 1), and there are an average of 7.4 introns per locus (11).

‡Blocks of up to 6,000 bp (excludes splice control regions).

§Blocks of up to 6,000 bp upstream or downstream from the coding sequence were analyzed.

¶Estimate based on ref. 8, but assuming 35,000 rather than 80,000 loci, calculated under the assumption that mice and rats diverged 13 million years ago (24) and have two generations per year (8).

divergence ≈15%). Mean constraint has dropped to levels close to zero by ≈4,000 bp from the coding sequence (Fig. 3*B*).

Contribution of Noncoding DNA to Overall Deleterious Mutation Rate. Under the assumption that there are 35,000 rodent genes (10, 11), we calculated the contributions of coding, intronic, and intergenic DNA to U (Table 4). In the set of loci analyzed, evolutionary constraint at amino acid sites calculated by a method as described (6) is 0.87 (SEM = 0.009), which is a typical value for rodent loci (33), and the contribution to U is 0.22. The overall estimate of U in noncoding DNA, summing over contributions from intronic and intergenic DNA, is also 0.22. This estimate for noncoding DNA is conservative for several reasons. (*i*) It does not include the contribution from constrained nucleotides outside the 6-kb 5′ and 3′ intergenic segments analyzed. This contribution is likely to be small, however, because ≈95% of well-characterized gene regulatory regions in murine intergenic regions are within 2 kb of promoters (11). (*ii*) The estimate will be too low if there are substantial selective constraints in FEIs. (*iii*) It does not include a contribution from indels. (*iv*) Our estimates of numbers of constrained nucleotides do not include sites under weak selection (with selection coefficients close to $1/N_e$). Such weakly selected mutations contribute to the mutation load (34, 35) and can have an appreciable probability of fixation, but the fraction of mutations with effects close to $1/N_e$ is relatively small for many reasonable distributions of selection coefficients.

In rodents, an overall estimate for U is ≈0.44 (Table 4). However, U is positively correlated with generation time (8), and U could be considerably higher for longer-lived taxa such as hominids. For example, an estimate for the mean level of constraint at amino acid sites in a sample of human and chimpanzee genes is 0.69 (33), and the generation time for hominids is ≈20 years (6). These estimates suggest that U for amino acid sites of protein-coding genes in hominids is ≈1.5 (8). If the proportion of deleterious mutations in noncoding DNA is similar among mammalian taxa, a genomic estimate for U (including point mutations in both coding and noncoding DNA) in hominids is therefore 3.0. Under a multiplicative model, the resulting mutation load (95%) is so high as to imply that nonmultiplicative effects of mutations are important in reducing the load in hominids.

The high frequency of deleterious mutations in intergenic DNA contrasts sharply with the low frequency of regulatory mutations associated with human Mendelian genetic diseases (≈1% of point mutations; ref. 36). This finding suggests that deleterious mutations in noncoding DNA are predominantly quantitative in nature and could be an important source of quantitative trait variation and of the burden of complex genetic disease in human populations. Human complex trait association mapping programs may therefore gain enhanced efficiency by concentrating markers in the regions of high constraint indicated by our study.

We thank Adam Eyre-Walker and Alex Kondrashov for helpful discussions and two reviewers for useful comments on an earlier version.

1. Crow, J. F. & Simmons, M. J. (1983) in *The Genetics and Biology of Drosophila*, eds. Ashburner, M., Carson, H. L. & Thompson, J. N. (Academic, London), Vol. 3C, pp. 1–35.
2. Crow, J. F. (2000) *Nat. Rev. Genet.* **1,** 40–47.
3. Muller, H. J. (1950) *Am. J. Hum. Genet.* **2,** 111–176.
4. Kondrashov, A. S. (1988) *Nature* **336,** 435–440.
5. Whitlock, M. C. (2002) *Genetics* **160,** 1191–1202.
6. Eyre-Walker, A. & Keightley, P. D. (1999) *Nature* **397,** 344–347.
7. Nachman, M. W. & Crowell, S. L. (2000) *Genetics* **156,** 297–304.
8. Keightley, P. D. & Eyre-Walker, A. (2000) *Science* **290,** 331–333.
9. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
10. International Human Genome Sequencing Consortium (2001) *Nature* **409,** 860–921.
11. Mouse Genome Sequencing Consortium (2002) *Nature* **420,** 520–562.
12. Clark, A. G. (2001) *Genome Res.* **11,** 1319–1320.
13. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.).
14. Kondrashov, A. S. & Crow, J. F. (1993) *Hum. Mutat.* **2,** 229–234.
15. Bergman, C. M. & Kreitman, M. (2001) *Genome Res.* **11,** 1335–1345.
16. Jareborg, N., Birney, E. & Durbin, R. (1999) *Genome Res.* **9,** 815–824.
17. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, F. A. & Kondrashov, A. S. (2001) *Trends Genet.* **17,** 373–376.
18. Hare, M. P. & Palumbi, S. R. (2003) *Mol. Biol. Evol.* **20,** 969–978.
19. Thorne, J. L., Kishino, H. & Felsenstein, J. (1991) *J. Mol. Evol.* **33,** 114–124.
20. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
21. Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. (2002) *Nat. Genet.* **31,** 415–418.
22. Calabrese, P. & Durrett, R. (2003) *Mol. Biol. Evol.* **20,** 715–725.
23. Li, W.-H., Wu, C.-I. & Luo, C.-C. (1984) *J. Mol. Evol.* **21,** 58–71.
24. Jaeger, J. J., Tong, H. & Denys, C. (1986) *C. R. Acad. Sci. Ser. II* **302,** 917–922.
25. Bird, A. (1986) *Nature* **321,** 209–213.
26. Rice, P., Longden, I. & Bleasby, A. (2000) *Trends Genet.* **16,** 276–277.
27. Antequera, F. & Bird, A. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 11995–11999.
28. Smith, N. G. C. & Hurst, L. D. (1999) *Genetics* **152,** 661–673.
29. Chen, F.-C. & Li, W.-H. (2001) *Am. J. Hum. Genet.* **68,** 444–456.
30. Hellmann, I., Zollner, S., Enard W., Ebersberger, I., Nickel, B. & Paabo, S. (2003) *Genome Res.* **13,** 831–837.
31. Sharp, P. A. (1994) *Cell* **77,** 805–815.
32. Majewski, J. & Ott, J. (2002) *Genome Res.* **12,** 1827–1836.
33. Eyre-Walker, A., Keightley, P. D., Smith, N. G. C. & Gaffney, D. (2002) *Mol. Biol. Evol.* **19,** 2142–2149.
34. Ohta, T. (1973) *Nature* **246,** 96–98.
35. Kondrashov, A. S. (1995) *J. Theor. Biol.* **175,** 583–594.
36. McKusick, V. A. (1998) *Mendelian Inheritance in Man: Catalogs of Human Genes and Genetic Disorders* (Johns Hopkins Univ. Press, Baltimore), 12th Ed.

8 Bennett, D.C. and Lamoreux, M.L. (2003) The color loci of mice – a genetic century. *Pigment Cell Res.* 16, 333–344
9 Rees, J.L. (2003) Genetics of hair and skin color. *Annu. Rev. Genet.* 37, 67–90
10 Eiberg, H. and Mohr, J. (1987) Major genes of eye color and hair color linked to LU and SE. *Clin. Genet.* 31, 186–191
11 Eiberg, H. and Mohr, J. (1996) Assignment of genes coding for brown eye colour (*BEY2*) and brown hair colour (*HCL3*) on chromosome 15q. *Eur. J. Hum. Genet.* 4, 237–241
12 Zhu, G. *et al.* (2004) A genome scan for eye colour in 502 twin familes: most variation is due to a QTL on chromosome 15q. *Twin Res.* 7, 197–210
13 Lee, S.T. *et al.* (1995) Organization and sequence of the human *P* gene and identification of a new family of transport proteins. *Genomics* 26, 354–363
14 Brilliant, M.H. (2001) The mouse *p* (pink-eyed dilution) and human *P* genes, oculocutaneous albinism type 2 (*OCA2*), and melanosomal pH. *Pigment Cell Res.* 14, 86–93
15 Rebbeck, T.R. *et al.* (2002) P gene as an inherited biomarker of human eye color. *Cancer Epidemiol. Biomarkers Prev.* 11, 782–784
16 Duffy, D.L. *et al.* (2004) Interactive effects of *MC1R* and *OCA2* on melanoma risk phenotypes. *Hum. Mol. Genet.* 13, 447–461
17 Kanetsky, P.A. *et al.* (2002) A polymorphism in the agouti signaling protein gene is associated with human pigmentation. *Am. J. Hum. Genet.* 70, 770–775
18 Frudakis, T. *et al.* (2003) Sequences associated with human iris pigmentation. *Genetics* 165, 2071–2083
19 Shriver, M.D. *et al.* (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum. Genet.* 112, 387–399
20 Chakraborty, R. and Weiss, K.M. (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. U. S. A.* 85, 9119–9123
21 McKeigue, P.M. *et al.* (2000) Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann. Hum. Genet.* 64, 171–186
22 McKeigue, P.M. (2000) Multipoint admixture mapping. *Genet. Epidemiol.* 19, 464–467
23 Swank, R.T. *et al.* (2000) Abnormal vesicular trafficking in mouse models of Hermansky-Pudlak syndrome. *Pigment Cell Res.* 13 (Suppl. 8), 59–67
24 Larsson, M. *et al.* (2003) Importance of genetic effects for characteristics of the human iris. *Twin Res.* 6, 192–200
25 Stjernschantz, J.W. *et al.* (2002) Mechanism and clinical significance of prostaglandin-induced iris pigmentation. *Surv. Ophthalmol.* 47 (Suppl. 1), S162–S175
26 Albert, D.M. *et al.* (2003) Iris melanocyte numbers in Asian, African-American, and Caucasian irides. *Trans. Am. Ophthalmol. Soc.* 101, 217–222
27 Ito, S. (2003) The IFPCS presidential lecture: a chemist's view of melanogenesis. *Pigment Cell Res.* 16, 230–236

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.06.010

Unexpected conserved non-coding DNA blocks in mammals

Daniel J. Gaffney and Peter D. Keightley

Ashworth Laboratories, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom

The significance of non-coding DNA is a longstanding riddle in the study of molecular evolution. Using a comparative genomics approach, Dermitzakis and colleagues have recently shown that at least some non-coding sequence, frequently ignored as meaningless noise, might bear the signature of natural selection. If functional, it could mark a turning point in the way we think about the evolution of the genome.

Few genomic features are more puzzling than the vast amounts of apparently functionless non-coding DNA that make up the greater proportion of human, mouse and many other eukaryotic genomes. However, although the view of non-coding sequence as genomic debris has been widespread, recent results by Dermitzakis and colleagues [1–3] offers a fascinating hint that a significant proportion can retain a function that, for the moment, remains a mystery.

For much of the past 50 years, the functional genome has been viewed as one that codes for protein and, until recently, most evolutionary studies of DNA sequences have focused almost entirely on this translated fraction, which we now think accounts for as little as 1–2% of both human and mouse DNA [4,5]. Many theories of the origin of non-coding DNA are founded on the perception that the bulk of such sequence is meaningless [6] and invoke random processes of accumulation of this 'junk', for example, the action of 'selfish' self-replicating elements [7]. Whole genome sequencing has, to some extent, borne these views out. Approximately 40% of mouse and human genomes are composed of the repetitive signatures that characterize past insertion of such retroelements [4,5]. Indeed, ~20% of the entire mouse genome appears to have originated via the activity of a single class of element, the long interspered elements (LINEs) [5]. However, excluding repetitive DNA sequence still leaves enormous quantities of non-coding sequence that we know little about. One of the most intriguing suggestions arising from the comparison of human and mouse genomes is that protein-coding sequences only account for approximately a fifth of the total amount of each species' genome that is subject to purifying selection [5]. The implication is that relatively large amounts of non-coding DNA are functional and it is clear, therefore, that the elucidation of potential functions (or otherwise) of non-coding DNA is a primary challenge in evolutionary genomics.

Corresponding author: Daniel J. Gaffney (Daniel.Gaffney@ed.ac.uk).

Development of comparative analyses

One powerful approach to address this challenge implements cross-species comparison of syntenic genomic regions with the aim of identifying potentially functional sequences. The underlying assumption, rooted in the neutral theory, is that conservation, above that expected given phylogenetic distance, implies selective constraint and, thus, function. Such 'phylogenetic footprinting' is not a new concept. Early comparative studies [8] were expanded by Duret *et al.* [9], who uncovered surprisingly strong conservation of flanking and untranslated regions (UTRs) of orthologous genes among widely diverged vertebrate groups. A similar approach on a larger scale by Koop *et al.* [10], comparing 100 kb of syntenic sequence surrounding T-cell receptor loci in mouse and human, revealed that conserved non-coding DNA could be found deep within intergenic regions in addition to sequence that is proximate to genes.

More recent human–mouse comparisons have revealed mosaics of constrained and randomly drifting sequence in intronic [11] and intergenic [12] DNA. The conserved blocks of non-coding sequence of which such mosaics consist have, in many cases, been found to correspond to exonic or regulatory regions [13,14]. Such pairwise comparisons are useful but can detect only a fraction of extant conservation and, as a result, multi-species phylogenies are becoming more commonly used. Analysis of sequence orthologous to a region of human chromosome 7 across 12 species by Thomas *et al.* [15] demonstrated the limitations of a single species pair in detecting the true pattern of non-coding conservation. High-density oligonucleotide arrays have also been employed recently in cross-species comparisons over entire chromosomes [16], revealing large quantities of conserved non-coding blocks that are not related to known exons. It is clear that large-scale comparative sequence analysis is rapidly becoming a useful tool to identify functionality within non-coding sequence [17,18].

Conserved non-genic sequences: functional non-coding DNA?

The work of Dermitzakis and colleagues [1] supports these previous findings and suggests that extensive conservation of non-coding DNA is not just a feature of a few mammalian species. Their initial study compared 33.5 Mb from the long arm of human chromosome 21 with syntenic sequence in mouse. The alignment of these regions revealed a high frequency of well-conserved, ungapped sequences located primarily in the Giemsa-dark, gene-poor region of chromosome 21 (Figures 1,2) a result that is supported by one previous study [16]. Initially, they removed those sequences containing existing, annotated exons and pseudogenes, which accounted for 1229 of the blocks sampled. However, the majority (2262) of the remaining conserved non-genic (CNGs) sequences appear not to match any of the ~230 known genes on chromosome 21. CNGs appear to be, on average, both smaller (~150 bp in length compared with the average chromosome 21 exon length of 270 bp) and more than twice as numerous as known exons. If functional, these sequences could include *cis*-regulatory elements, undiscovered protein-coding

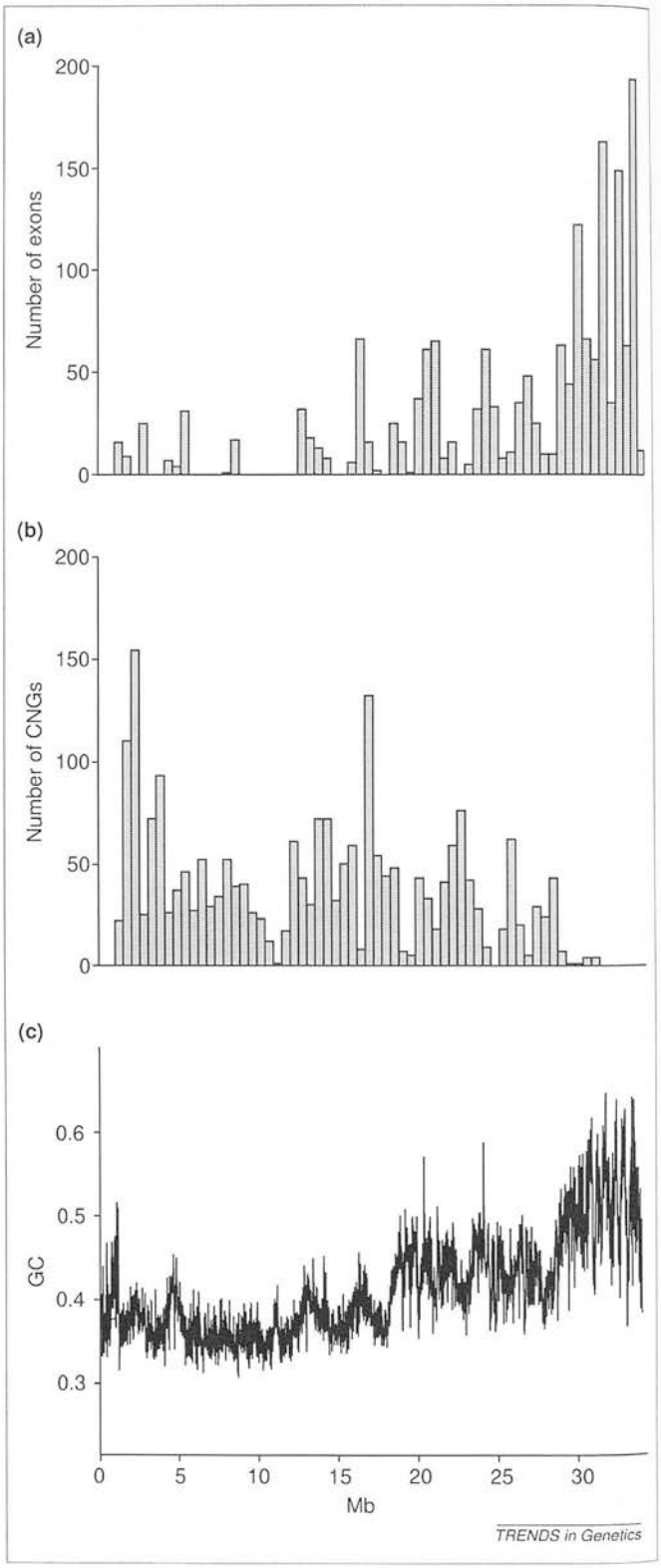

Figure 1. The frequency of known exons **(a)** and conserved non-genic sequences (CNGs) **(b)** varies nonrandomly with respect to local GC content **(c)** across the long arm of human chromosome 21, starting at the centromere. Exons tend to be located in the GC rich distal end, whereas CNGs are predominantly located in the AT rich region proximal to the centromere. The average local GC content was calculated in non-overlapping 15-kb windows.

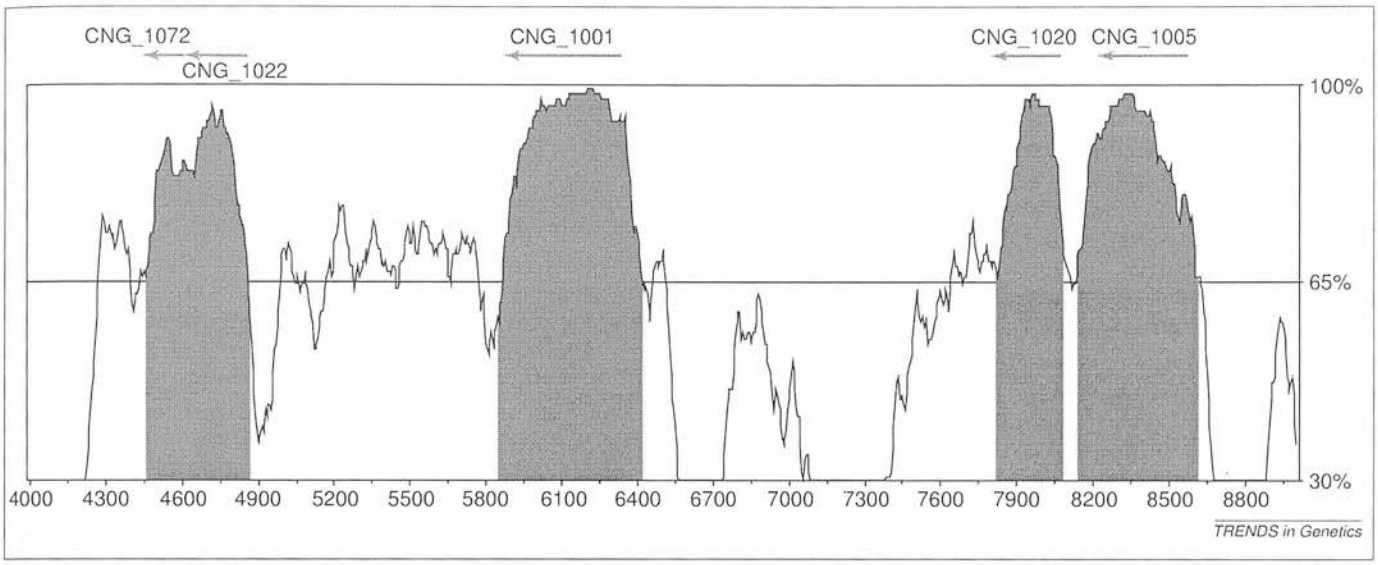

Figure 2. Conserved non-genic sequences (CNGs) can be extremely highly conserved when compared with background sequence divergence, which could suggest a selectively preserved function. This mVISTA [28,29] plot shows sequence similarity in an alignment of a 5-kb region of human chromosome 21 containing four CNGs (annotated in the supplementary materials of Ref. [1]) with a syntenic region located on mouse chromosome 16. Each CNG is annotated in red and has an average divergence of <15% compared with the average human–mouse neutral divergence of 35%. mVISTA software is also available on a public web server at http://gsd.lbl.gov/vista/index.shtml.

exons, non-coding RNA genes (ncRNAs) and undescribed functional non-coding sequences. Alternatively, conservation could simply result from nonfunctional mutational 'coldspots'.

Initially, the authors combined computational and experimental analyses to investigate the protein-coding potential of their data. Although similarity-based methods are often the most reliable methods of gene recognition, they are, by definition, limited by the availability of known protein-coding regions in related species. Instead, CNGs were compared with those exons predicted by the GrailEXP (http://compbio.ornl.gov/grailexp/), Pro-Gen (http://www.anchorgen.com/pro_gen/pro_gen.html) packages and GenomeScan (http://genes.mit.edu/genomescan/). Such prediction algorithms generally operate on the principal that various compositional features are conserved across genes and can be used to predict probable coding structures in an unknown sequence. CNGs were also compared with existing human and mouse EST databases for similarities to existing cDNAs. These predictive approaches almost uniformly returned low numbers of candidate protein-coding genes. The authors also estimated the non-synonymous versus the synonymous (K_a/K_s) substitution ratio in the six potential reading frames of all CNGs; this is typically <0.3 for most functional human–mouse orthologues [5]. Blocks within 100 kb of each other with low K_a/K_s ratios were identified as potential exon pairs. However, reverse transcription amplification (RT–PCR) of such 'adjacent blocks' indicated that only a small fraction produce mature mRNAs.

Combining the results of these analyses appeared to indicate that the coding potential of the majority of CNGs is small. Moreover, the distances between consecutive substitutions within CNGs appear randomly distributed compared with the highly distinctive 'triplet' pattern observed in coding sequence, which results from the elevation of substitution rates at codon third positions. Taken by itself, the apparent lack of protein-coding sequence in the data is unexpected. If we assume that most exons on human chromosome 21 are conserved between human and mouse, then it seems reasonable that most exons in this region, known and unknown, were part of the initial 3491 blocks that were detected. The removal of known exons and pseudogenes implies that only unknown exons, along with conserved non-coding sequence, could possibly have remained in the dataset. The failure to detect virtually any coding sequence in these remaining conserved blocks leaves two possible conclusions: (i) few unknown exons that are conserved between human and mouse exist within the sample region; or (ii) the variety of prediction methods employed have failed to detect unannotated exons that remain within the data.

Defining a null hypothesis: the trouble with comparative genomics

Although the results of Dermitzakis *et al.* would indicate that the CNGs they observe are not exonic, this does not consider the possibility that the conservation observed is merely a product of low divergence. The authors addressed this by attempting to amplify a subset of CNGs in rabbit [1] and a variety of more and less diverged mammalian outgroup species [2]. The rationale behind this is straightforward. Conservation in more widely diverged taxa provides further evidence that blocks are preserved as the result of selective optimisation of function, given that point mutations will have had more time to randomise neutral sequence between functional modules in more distant outgroups. Perhaps their most startling result shows that a fraction of CNGs is highly conserved in a wide variety of mammals. The level of conservation exceeds that observed in both known ncRNAs and coding sequence. Furthermore, comparison of the entire CNG dataset with the canine genome returned a high frequency of reciprocal best hits indicating that appreciable conservation across species might not be confined to the amplified subset. This

means that, if functional, at least some CNGs are sufficiently important to be retained, almost unchanged, across multiple, divergent evolutionary paths.

The possible functions of CNGs remain unclear. Conservatively, their numbers are estimated to be almost twice that of the predicted protein-coding sequences, and a large fraction of CNGs might be functional. However, although sequence conservation is often a signature of the operation of selective forces, we should proceed with caution before accepting conservation *ipso facto* as proof of function. The fact remains that one of the most difficult objectives in any attempt to locate functional blocks via comparative analysis is the definition of a robust null hypothesis. A comprehensive understanding of the pattern of mutational variability across the genome is fundamental to the success of comparative genomics. Our knowledge of such variability, however, is still incomplete although this could explain at least some observed conservation [19]. One approach that can, at least partially, elucidate the pattern of evolution we expect under neutrality is a comparison of rates of change within putatively functional sequences that have elements adjacent, such as ancestral repeats, which we can be relatively sure are evolving free of selection. We have used this approach in a recent study [20] to examine patterns of constraint in rodent non-coding sequence under the assumption that, excluding various potentially important regions such as donor and acceptor splice sites, intronic sequence is evolving neutrally. Our results support those of Dermitzakis *et al.* in that we also find comparatively large quantities of conserved non-coding sequence. However, it is important to note the conservation we observed is relative to the local, presumably neutral substitution rate, within introns and is thus, not directly comparable with that observed by Dermitzakis *et al.* Our analysis reveals a steady decrease in conservation with increasing distance upstream and downstream of a protein-coding sequence to approximately zero within 4 kb of most genes. By contrast, CNGs appear markedly disassociated with annotated exons: we infer a negative and highly significant correlation between the numbers of each within intervals along the chromosome (Pearson $r = -0.47$; $P < 0.001$). The large median distance between known coding sequence and CNGs suggests that there is unlikely to be much overlap between our datasets, and, indeed, recent results have indicated that conservation of CNGs is independent of their position relative to known genes [3].

The comparatively low GC content of CNGs (~38%) appears to reflect their genomic position, with the majority of CNGs located in the GC-poor proximal half of chromosome 21 (Figure 1). GC-poor regions are expected to be characterized by lower substitution rates than those in GC-rich sequence because of the lower frequency of hypermutable CpG dinucleotides. Although this could explain at least some of the unexpected conservation of CNGs between species, we estimate that both local GC content and CpG dinucleotide frequency explain little of the variation in human–mouse divergence. In addition, such causality would also require the conservation of base composition between species that for the human–mouse–rabbit comparison appears not to be the case.

If CNGs are not exonic, are they RNA genes?

Of the working hypotheses that were originally suggested, we can reasonably conclude that the hidden or unannotated exon theory has been the most convincingly eliminated. It seems unlikely that such a comparatively large number of otherwise unrelated sequences would fail to produce an appreciable frequency of predicted genes, spliced transcripts or matches with known cDNAs if they were exonic. The case for several ncRNAs being present within the CNG data is perhaps more convincing. BLAST comparison of CNGs with data from a transcriptional study of human chromosome 21 [21] indicated that approximately a fifth matched sequences that are known to be transcribed. On the basis of substitution pattern within blocks, ncRNA prediction software returned a comparable number of probable ncRNAs within the data. It would be interesting to investigate the degree of overlap between matched oligonucleotides and those predicted RNA gene models and to verify some of the predicted RNA gene models experimentally. Dermitzakis and colleagues also attempted a fine-scale analysis of substitutional patterns, testing for significant differences in clustering of variable sites and substitutional asymmetries between CNGs, coding sequence and ncRNAs. Although the distribution of variable sites appears to distinguish CNGs from protein-coding sequence, the results are uncertain in the case of ncRNAs, where the patterns of evolution appear to be highly variable. Furthermore, the estimates of substitution rate from the admittedly small number of known ncRNAs indicates that they are more highly conserved than coding sequences obtained by Dermitzakis *et al.* Although this could indicate a conservation bias in known ncRNAs it also reflects how little we know about the evolutionary significance of such genes.

It is important to note that, despite the high degree of similarity between some CNGs across many species, identity by itself does not confirm the orthology of CNGs among groups, although the probability of sequences of such high similarity being non-orthologous is small. We can be almost certain of the orthology of CNGs in the human–mouse comparisons, which were based on the criteria of synteny. The same can not be said of those comparisons with rabbit or more distant species. This might constitute a weakness of the approach of Dermitzakis *et al.* and contrasts with the targeted sequencing protocol that was implemented by Thomas and colleagues [15] where a region known to be orthologous to a segment of human chromosome 7 was amplified in multiple species. It is notable, however, that these authors' conclusions do not depart greatly from those of Dermitzakis *et al.*, in that they also observed a high degree of sequence conservation in presumed functional 'deserts'.

Also the subset of 220 CNGs that was selected for PCR might be slightly biased towards higher conservation, presumably to increase the frequency of PCR success, compared with the original dataset of 2262. The selection of this subset is of crucial importance in determining whether high conservation is a feature of all CNGs. The high frequency of reciprocal best hits in the dog genome for the entire CNG dataset would support the conclusion that most blocks are conserved across species. Confirmation of further conserved blocks in other species is required, however, before we will know the true prevalence of CNGs that have evidence of functionality.

Concluding remarks

These caveats notwithstanding, Dermitzakis *et al.* have shown that evolutionary conservation of non-coding blocks on chromosome 21 extends across the entire mammalian lineage, a result that, in itself, is fascinating. However, although a comparative study can identify candidate functional regions, characterization must typically be sought via experimental assay. An approach employing these two steps to great effect in a recent study [22] identified long-range regulatory elements of several important cytokines. Other experimental research is now beginning to shed light on the possible functions that CNGs could perform.

Functional assays have provided evidence of regulatory motifs that operate at great distances from their associated genes [23]. In addition, substantial transcriptional activity has recently been reported outside of annotated regions along chromosome 21, although the proportions that are due to coding versus non-coding regions are difficult to ascertain [24]. Furthermore, the distribution of the binding sites of well-characterized human transcription factors on chromosome 21 suggests that many are located outside defined promoter regions [25] and there appears to be evidence of the regulatory function of a small number of CNGs on chromosome 21 located around the single-minded homolog 2 (*SIM2*) transcription factor [26]. Whether such examples are representative of *cis*-regulatory regions as a whole and whether the pattern of conservation of CNGs is indicative of such function remains uncertain. One interesting consequence of the GC-poor location of most CNGs is that the genes they are proximate to tend to have large introns and often require complex regulation. By contrast, housekeeping genes, which require little transcriptional regulation, tend to be situated in GC-rich (and thus CNG-poor) regions. Thus, there is again a suggestion that, if functional, some CNGs might have a role in long-distance expression control. In addition, conserved non-coding sequence might not only function in gene regulation but also might have a structural role, for example, as matrix scaffold attachment regions [27].

It is the 'second-steps' of experimental characterization, such as the studies described previously, which have perhaps the most intriguing possibilities for further work. If they are functional, the conserved regions being uncovered using the comparative method could finally begin to unravel the secrets of non-coding DNA.

Acknowledgements

We thank Mark Blaxter, Dan Halligan and the anonymous referees for valuable comments.

References

1 Dermitzakis, E.T. *et al.* (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420, 578–582
2 Dermitzakis, E.T. *et al.* (2003) Evolutionary discrimination of mammalian conserved non-genic sequences. *Science* 302, 1033–1035
3 Dermitzakis, E. *et al.* (2004) Comparison of human chromosome 21 conserved non-genic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* 14, 852–859
4 International Human Genome Sequencing Consortium, (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
5 International Mouse Genome Sequencing Consortium, (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
6 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404
7 Orgel, L. and Crick, F. (1980) Selfish DNA: The ultimate parasite. *Nature* 284, 604–607
8 Yaffe, D. *et al.* (1985) Highly conserved sequences in the 3′ untranslated region of messenger-RNAs coding for homologous proteins in distantly related species. *Nucleic Acids Res.* 13, 3723–3737
9 Duret, L. *et al.* (1993) Strong conservation of noncoding sequences during vertebrates evolution – potential involvement in posttranscriptional regulation of gene-expression. *Nucleic Acids Res.* 21, 2315–2322
10 Koop, B.F. and Hood, L. (1994) Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* 7, 48–53
11 Jareborg, N. *et al.* (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* 9, 815–824
12 Shabalina, S.A. *et al.* (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* 17, 373–376
13 Ansari-Lari, M.A. *et al.* (1998) Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res.* 8, 29–40
14 Oeltjen, J.C. *et al.* (1997) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* 7, 315–329
15 Thomas, J.W. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788–793
16 Frazer, K.A. *et al.* (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* 11, 1651–1659
17 Dubchak, I. and Frazer, K. (2003) Multi-species sequence comparison: the next frontier in genome annotation. *Genome Biol.* 4, 1221–1226
18 Elnitski, L. *et al.* (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res.* 13, 64–72
19 Clark, A.G. (2001) The search for meaning in noncoding DNA. *Genome Res.* 11, 1319–1320
20 Keightley, P.D. and Gaffney, D.J. (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13402–13406
21 Kapranov, P. *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919
22 Loots, G.G. *et al.* (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288, 136–140
23 Nobrega, M.A. *et al.* (2003) Scanning human gene deserts for long-range enhancers. *Science* 302, 413
24 Kampa, D. *et al.* (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* 14, 331–342
25 Cawley, S. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116, 499–509
26 Frazer, K. *et al.* (2004) Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* 14, 367–372
27 Glazko, G.V. *et al.* (2003) A significant fraction of conserved noncoding

DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.* 19, 119–124
28 Mayor, C. *et al.* (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16, 1046–1047
29 Dubchak, I. *et al.* (2000) Active conservation of noncoding sequences revealed by threeway species comparisons. *Genome Res.* 10, 1304–1306

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.06.011

Antibody class switching: uncoupling S region accessibility from transcription

Denise A. Kaminski and Janet Stavnezer

Department of Molecular Genetics and Microbiology, Program in Immunology and Virology, University of Massachusetts Medical School, Worcester, MA 01655, USA

Immunogloblin class switch recombination (CSR) is a regulated process that changes antibody effector functions. Recently, Nambu *et al.* showed that histone acetylation is induced at switch (S) regions undergoing CSR; however, histone acetylation without accompanying S region transcription is insufficient to attract activation-induced cytidine deaminase (AID), which is required for CSR. They also show that AID can associate with RNA polymerase II. These results support the model that germline transcripts are required to form single-stranded DNA, the AID substrate and further suggest that AID is recruited to S regions by the transcriptional machinery.

Activation of antibody-producing B lymphocytes (B cells) by antigen and co-stimulatory signals results in class switch recombination (CSR). CSR enables B cells to change the antibody constant (C) region, enhancing the ability of the antibody to eliminate pathogens, while maintaining the same antigen-binding variable region (Figure 1a). Following activation of mouse B cells, the Cμ gene encoding the heavy chain C region of the initial IgM class can be replaced by any of the downstream C genes, resulting in expression of IgG3, IgG1, IgG2b, IgG2a, IgE or IgA (Figure 1b). CSR occurs by deletional recombination between repetitive G-rich switch (S) region sequences located upstream of each C gene except Cδ. The intervening DNA is then detected as an extrachromosomal switch circle [1]. CSR requires the B-cell-specific enzyme, activation-induced cytidine deaminase [AID (see Glossary)] [2,3]. Owing to its homology with an mRNA-editing cytidine deaminase, AID was first proposed to initiate CSR indirectly by editing an mRNA to create a novel endonuclease. However, accumulating evidence is consistent with AID acting directly on the DNA of the Ig S REGION, converting dC to dU residues, thereby initiating DNA repair processes that result in the DNA breaks required for CSR [4–8]. How AID is targeted to S regions has remained unclear and is the focus of a recent report by Nambu *et al.* [9].

Corresponding author: Janet Stavnezer (Janet.Stavnezer@umassmed.edu).
Available online 19 June 2004

S region accessibility during CSR

It has been proposed that S region susceptibility to recombination is caused by the obligatory transcription of S regions in their germline (GL) configuration [10–13]. This transcription initiates upstream (5′) of the target S regions and continues through the C region exons (Figure 1b). GERMLINE TRANSCRIPTION is induced specifically by the same stimuli that induce CSR to each specific S region, namely B-cell activators, such as, lipopolysaccharide (LPS) together with cytokines [10]. The germline transcripts (GLTs) must also be spliced, however, suggesting that the function of GL transcription is not simply to induce accessibility of S region chromatin [10,12,13]. This hypothesis has been recently supported and extended by results from Nambu *et al.* [9], which suggest that AID association with S region DNA requires GL transcription and not simply chromatin accessibility.

To examine the accessibility of chromatin associated with specific S regions during CSR, Nambu *et al.* [9] used chromatin immunoprecipitation (ChIP) to assay HISTONE ACETYLATION, a modification known to occur in chromatin associated with actively transcribed genes [14]. Mouse splenic B cells were treated with LPS and interlukin 4 (IL-4), which induces GLTs initiating upstream of the γ1 and ε switch regions, and subsequently CSR to IgG1 and to

Glossary

AID: activation-induced cytidine deaminase; a B-cell-specific enzyme that is required for class switch recombination (CSR) and for somatic hypermutation of antibody variable region genes.
S region: 2–10-kb DNA segments located upstream of heavy chain constant (C) genes where switch recombination occurs. S regions consist of tandem repeats of 20–80-bp consensus sequences and contain frequent GAGCT, GGGGT and GGGCT elements.
Germline (GL) transcription: RNA synthesis that initiates upstream of switch (S) regions and is necessary for CSR. Although the GL transcripts are polyadenylated and spliced, no protein products have been detected.
R-loop: a non-coding nucleic acid structure consisting of RNA that forms a duplex with one strand of DNA, leaving the non-coding DNA strand single stranded.
UNG: the uracil DNA glycosylase required for normal levels of CSR. This might be due to its ability to excise dU bases that result from AID activity in genomic DNA.
Histone acetylation: modification of histone proteins by the addition of acetyl groups; this modification occurs most commonly on lysine residues of the N-terminal tails of histones H3 and H4. It is associated with actively transcribed genes.

The scale of mutational variation in the murid genome

Daniel J. Gaffney[1] and Peter D. Keightley

Institute of Evolutionary Biology, Ashworth Laboratories, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Mutation rates vary across mammalian genomes, but little is known about the scale over which this variation occurs. Knowledge of the magnitude and scale of mutational variation is required to understand the processes that drive mutation, and is essential in formulating a robust null hypothesis for comparative genomics studies. Here we estimate the scale of mutational variation in the murid genome by calculating the spatial autocorrelation of nucleotide substitution rates in ancestral repeats. Such transposable elements are good candidates for neutrally evolving sequence and therefore well suited for the study of mutation rate variation. We find that the autocorrelation coefficient decays to a value close to zero by ~15 Mb, with little apparent variation in mutation rate under 100 kb. We conclude that the primary scale over which mutation rates vary is subchromosomal. Furthermore, our analysis shows that within-chromosome mutational variability exceeds variation among chromosomes by approximately one order of magnitude. Thus, differences in mutation rate between different regions of the same chromosome frequently exceed differences both between whole autosomes and between autosomes and the X-chromosome. Our results indicate that factors other than the time spent in the male germ line are important in driving mutation rates. This raises questions about the biological mechanism(s) that produce new mutations and has implications for the study of male-driven evolution.

[Supplemental material is available online at www.genome.org.]

Much evidence now suggests that the point mutation rate varies considerably across the mammalian genome. Studies of nucleotide substitution rates at synonymous sites (Wolfe et al. 1989; Matassi et al. 1999; Malcom et al. 2003; Chuang and Li 2004), within long alignments of primate intergenic sequence (Chen et al. 2001; Ebersberger et al. 2002; Silva and Kondrashov 2002; Smith et al. 2002) and mammalian repetitive sequence (Waterston et al. 2002; Hardison et al. 2003), have revealed considerably more variation in the substitution rate than expected by chance. This is of interest because substantial mutational variability could seriously reduce the effectiveness of comparative methods to locate putatively functional regions within noncoding DNA. The efficiency of identification of such regions could be improved if we knew a priori which regions are expected to be evolving more slowly.

The regional mutation hypothesis proposes that different regions of the vertebrate genome are diverging at substantially different rates (Filipski 1988). Previous studies have provided evidence that mutation rates vary substantially between chromosomes (Wolfe et al. 1989; Malcom et al. 2003; Makova et al. 2004). Particularly notable is the apparent reduction in the rate of point (McVean and Hurst 1997; Ebersberger et al. 2002; Waterston et al. 2002) and indel substitution (Makova et al. 2004) on the X-chromosome. This reduction has been suggested to reflect the primarily male origin of most mutations, although the evidence on this point is inconsistent (McVean and Hurst 1997; Lercher et al. 2001). In addition, there is evidence that significant variation in the mutation rate also occurs along the length of a chromosome (Wolfe et al. 1989; Chuang and Li 2004). Although mutational variation has been studied at these two levels, an unresolved problem is the relative importance of chromosome number and position within a chromosome in determining the underlying mutation rate. Of particular relevance to this question is the scale of local similarity of mutation rates. If the domain or "unit" of mutational variation is considerably smaller than a chromosome and substantial interdomain variability exists, this would suggest that position within a chromosome is a more important factor in determining neutral mutation rate. This conclusion is reversed if local similarity extends across entire chromosomes.

One of the first studies to address the issue of local similarity of evolutionary rates compared estimates of the synonymous divergence (K_S) from human–mouse gene orthologs within 1 cM of each other, and concluded that there is evidence for the existence of "evolutionary rate units" between which substantial variation exists (Matassi et al. 1999). Lercher et al. (2001) extended this analysis to a larger data set and found that significant similarity of K_S extends from 1 cM to entire chromosomes in a human–rodent comparison. Although it may seem unexpected that mutation rates would remain approximately constant across entire chromosomes, this situation does appear to exist in yeast (Chin et al. 2005). Such a large scale of similarity would seem to reject a substantial role for within-chromosomal mutational heterogeneity and apparently suggests that the majority of mutational variation occurs between chromosomes. However, more recent work has suggested that synteny blocks (i.e., regions for which gene order has been conserved between species) may represent a more meaningful "unit" than whole chromosomes (Malcom et al. 2003; Webster et al. 2004). Malcom et al. (2003) found that although a weak effect of chromosomal number is evident in both human–mouse and mouse–rat comparisons, this is confounded by substantial within-chromosome variation. These au-

[1]**Corresponding author.**
E-mail Daniel.Gaffney@ed.ac.uk; fax 44-131-6506564.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3895005. Article published online before print in July 2005.

thors indicate that differences between synteny blocks on the same chromosome outweigh those observed between chromosomes. Additional support for a subchromosomal mutational scale comes from Chuang and Li (2004), who use a human–mouse comparison to show that local similarity in mutation rates extends to ~10 Mb. The relevance of a chromosome as an evolutionarily distinct entity is uncertain, however, particularly between highly diverged species such as human and mouse, for which genome sequencing projects have revealed many large-scale rearrangements (Nadeau and Taylor 1984; Hudson et al. 2001; Waterston et al. 2002).

Many of the above studies have used synonymous substitution rates to examine patterns of mutational variation. However, synonymous sites comprise a small fraction of most mammalian genomes and may misrepresent mutational processes outside of coding sequence. In addition, the importance of sequence context effects, in particular CpG hypermutability, is becoming increasingly apparent (Arndt et al. 2003). Given that the majority of sites both 5′ and 3′ of mammalian fourfold degenerate synonymous sites are under strong purifying selection, this may introduce bias in the estimation of K_S. For example, strong selective preservation of a C that is 5′ to a fourfold synonymous site may serve to elevate the observed substitution rate. Furthermore, there is now some evidence that selection, perhaps related to mRNA splice efficiency or mRNA stability, may be operating at some mammalian synonymous sites (Eyre-Walker 1999; Keightley and Gaffney 2003; Chamary and Hurst 2004; Willie and Majewski 2004; Keightley et al. 2005).

For these reasons, it is desirable to investigate mutational variation outside of coding sequence. Some authors have sought to address this by using long human–chimpanzee alignments of intergenic sequence (Ebersberger et al. 2002; Smith et al. 2002; Webster et al. 2004). Webster et al. (2004) estimated the extent of local similarity using substitution rates at ancestral repeat (AR), intronic, and intergenic sites from a human–chimp alignment of 14 Mb from human Chromosome 7. Their results indicate that the most significant local similarity of mutation rates occurs at a scale of 1–2 Mb. However, they did not investigate the rate of decay of this local similarity. Furthermore, it is becoming increasingly apparent that some of the noncoding nonrepetitive portion of the mammalian genome, assumed to be neutral in the above studies, may be under selection (Waterston et al. 2002; Thomas et al. 2003; Bejerano et al. 2004). Smith et al. (2002) and Webster et al. (2004) argue that such selected regions should have little influence on substitutional variation in closely related species. However, minimally diverged species are more susceptible to the influence of ancient polymorphism in the last common ancestor, and selection in noncoding DNA does become relevant when considering alternative, more distantly related taxa, such as mouse and rat. Thus, in these species pairs, long, intergenic alignments are not ideal for the study of mutational variation. One alternative is to focus on the remnants of repetitive elements that were inserted in the last common ancestor (e.g., Waterston et al. 2002; Hardison et al. 2003) The use of these ancestral repeats is appealing because, of all classes of noncoding DNA, they are the most likely candidates for neutrality (Ellegren et al. 2003). Additionally, the large quantities of repetitive sequence allow for investigation of mutational variation on much finer scales than is possible just using K_S.

We therefore collected a data set of repetitive elements present in the last common mouse–rat ancestor. Using these data, we sought to address the following questions: (1) What is the scale of local similarity of rodent mutation rates? (2) At this scale, what is the ratio of between-chromosome to within-chromosome mutation rate variation? Answers to these questions are important to accurately quantify mutational variation and improve our understanding of the processes that may cause point mutation. Furthermore, information on the scale of mutational variation is important in establishing a robust null hypothesis for comparative genomics methods.

Results

We extracted and aligned a total of 55 Mb of repetitive sequence. This can be broken down into the following contributions from various classes of repetitive elements: 17.5 Mb of SINE, 13.0 Mb of LINE, 21.0 Mb of LTR, and 3.7 Mb of DNA elements. The proportions of aligned sequence derived from each repeat family appears approximately consistent across autosomes (Fig. 1). However, LINE elements appear to be significantly more prevalent on the X-chromosome ($P < 0.0001$) than the autosomes. This would suggest either that LINE elements have been more active on the X-chromosome or that the rate of deletion of LINEs is less than on the autosomes. There is some evidence to suggest that the former scenario is more likely, as it seems that some retrotransposing sequences preferentially target the X-chromosome (Khil et al. 2005). It may also be that LINEs play a role in X-chromosome inactivation (Bailey et al. 2000; Waterston et al. 2002).

Between-chromosome variation

We estimated the average chromosomal divergence at all sites and at sites not preceded by a C or followed by a G (non-CpG-prone sites) for each mouse chromosome (Fig. 2). Non-CpG-prone sites are the least likely to have been part of a hypermutable CpG dinucleotide, and therefore the least affected by potential covariation between nucleotide divergence and age of transposable element insertion (see Methods). We find that the X-chromosome is evolving more slowly at all sites than any of the autosomes, and we estimate a male-to-female mutation rate, α, of 1.5. This is slightly lower than previous estimates in rodents (Chang et al. 1994; Gibbs et al. 2004). Rates at non-CpG-prone sites are consistently lower than those estimated at all sites for all autosomes. This would suggest that rates at all sites are affected by the elevated mutation rates at CpG dinucleotides and the selection of non-CpG-prone sites goes some way to removing this effect. Interestingly, however, this situation is reversed on the

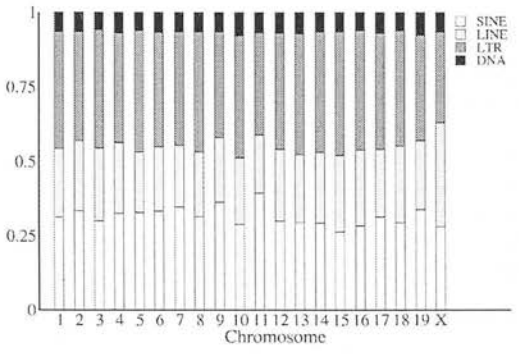

Figure 1. Proportion of total sequence per mouse chromosome contributed by each repeat class.

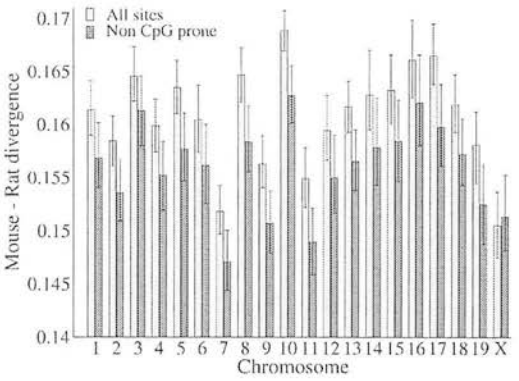

Figure 2. Estimated average nucleotide substitution rates at all sites and non-CpG-prone sites for each mouse chromosome. Bars show the 95% bootstrap confidence intervals.

X-chromosome, where substitution rates at non-CpG-prone sites are, in fact, marginally, although not significantly, higher than those estimated at all sites. This result appears to be roughly consistent within repeat families (Supplemental Table 1).

Scale of local similarity

We estimated the scale of local similarity of mutation rates using the autocorrelation of average substitution rates across a variety of block sizes. Figure 3 shows the autocorrelation of nucleotide substitution rates at all sites between blocks of 5 kb and 100 kb extending over intervals from 10 kb to 1 Mb and 200 kb to 20 Mb, respectively. Autocorrelation of rates across 5-kb blocks (Fig. 3A) remains highly significant compared to randomly permuted data across a distance of 1 Mb. There is minimal change in autocorrelation from 10 kb to 100 kb (Fig. 3A), suggesting that little variation in underlying mutation rate exists below 100 kb. The low magnitude of the correlation across 5-kb blocks reflects the relatively noisy estimates of substitution rates obtained from the small number of ancestral repeat sites (295 bp on average) within each block. In contrast, the number of sites within the average 100-kb block is approximately one order of magnitude larger than that in 5-kb blocks (2.3 kb on average), thus our estimate of the substitution rate is less noisy and the magnitude of autocorrelation is higher. Here, there is a slow decay of similarity in substitution rates extending to a distance of 10–15 Mb (Fig. 3B). It is important to note that autocorrelation in Figure 3A,B shows the same proportional change over the same distance. For example, autocorrelation across 5-kb blocks decays from ~0.078 to ~0.052 (a decrease of approximately one-third) over a distance of 1 Mb; autocorrelation across 100-kb blocks decays from ~0.445 to ~0.290 (again a decrease of approximately one-third) over the same distance.

The similarity of evolutionary rates between blocks within an interval of 0–15 Mb seems to be explained, in part, by the corresponding similarity of average GC content of adjacent blocks, since randomly permuting blocks within GC classes still produces a moderate signal of autocorrelation in the absence of local structure (Fig. 3C,D). This would suggest that local GC content, or one or more covariates of local GC content, influences neutral substitution rates in both repetitive and nonrepetitive DNA. However, this similarity does not seem to be a result of CpG hypermutability or compositional change, since our results were qualitatively similar when we estimated rates at non-CpG-prone sites or by counting A↔T and G↔C changes only (Supplemental Fig. 1).

We also estimated the partial autocorrelation of nucleotide substitution rates in both ancestral repeats and flanking sequence, averaged across 100-kb blocks (Fig. 4). Plots of partial autocorrelation coefficients suggest that all local similarity over distances >1 Mb can be explained by autocorrelations below 1 Mb. This suggests that the average "unit" of mutational variation is no larger than ~1 Mb. The results are similar in both repetitive and nonrepetitive sequence (Fig. 4A,B, respectively).

Within- and between-chromosome mutational variation

The data were initially fitted to two linear models, one including terms for fixed chromosomal and random regional effects, and the other including a chromosomal effect only. We estimated the magnitude of within-chromosome mutational variation as the variation between levels of the random regional effect in the former model. Model fit was assessed using Akaike's Information Criterion (AIC). It appears that all models that included "regional variation" effects provide a substantially better fit to the data than those including chromosome means alone (Fig. 5). This is clearly seen from the decrease in AIC (models with a better fit have a lower AIC) for models including a random regional effect. The AIC for Model 1 was −844,146.7 for ancestral repeats and −933,598.5 for flanking sequence data. Including blocks of 1 Mb as a random effect in the model, for example, decreases the AIC to −854,057.2 for the ancestral repeat data (Fig. 5) and to −945,433.5 for the flanking sequence data. This is evidence that significant regional variation in neutral mutation rate does, indeed, occur along the length of a chromosome. The most parsimonious model (as adjudged by the AIC) in our analysis includes a block size of 1 Mb as a random effect. At this scale the variation between blocks is approximately one order of magnitude greater than that observed between chromosomes. The between-chromosome variance was 2.28×10^{-5} for ancestral repeats and

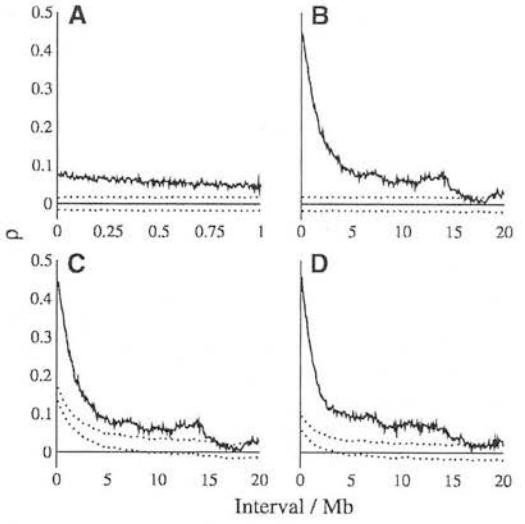

Figure 3. Autocorrelation of nucleotide substitution rates in ancestral repeats (*A*, B, C) and ancestral repeat flanking sequence (*D*) across 5-kb (*A*) and 100-kb (*B*, C, *D*) blocks. Substitution rates were estimated at all sites. Dotted lines show the upper and lower bounds of the 95% confidence interval of autocorrelation under the null hypothesis of no dependence of rates between blocks. Blocks were permuted randomly (*A*,*B*) and within common GC-content intervals (*C*,*D*).

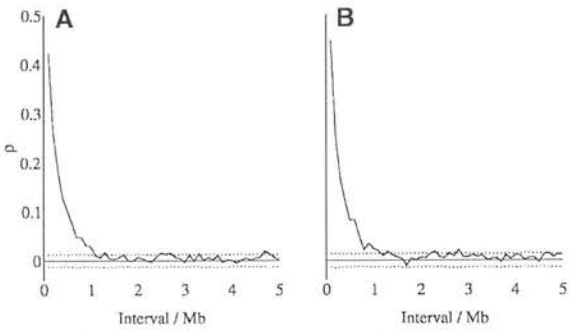

Figure 4. Partial autocorrelation of nucleotide substitution rates in ancestral repeats (*A*) and flanking sequences (*B*). Substitution rates are estimated for all sites. Dotted lines show the upper and lower bounds of the 95% confidence interval of partial autocorrelation under the null hypothesis of no dependence of rates between blocks.

7.71×10^{-7} for flanking sequence, whereas the between-block variance in the most parsimonious model is 2.06×10^{-4} for ancestral repeats and 9.53×10^{-5} for flanking sequence. While the substitution rates in ancestral repeats appear more variable than flanking nonrepetitive sequence, the difference in between- and within-chromosome mutational variation is striking in both categories of sites. Our results are consistent whether we consider rates at non-CpG-prone sites or by counting only A↔T and G↔C changes (Supplemental Figs. 2 and 3).

We also determined whether there were significant chromosome effects by comparing the mixed model (Model 2) with a model that includes a term for random regional effects only (Model 3). Regional effects of 1 Mb were included in both models. We analyzed four different data sets, consisting of nucleotide substitution rates in ancestral repeats and flanking sequence, including and excluding the X-chromosome. Our results indicate that Model 2 describes the data most parsimoniously in all cases (Table 1). We note, however, that the difference in AIC between Model 2 and Model 3 is far smaller (approximately two orders of magnitude) than that observed between Model 1 and Model 2. This would support our conclusion that although there exist small but detectable chromosomal effects on nucleotide substitution rates, they are far outweighed by subchromosomal regional variation. Differences in AIC between Model 2 and Model 3 drop when the X-chromosome is excluded.

We investigated the efficiency of our approach by analyzing simulated data (Supplemental material). Results of this analysis indicate that when regional effects are absent, Model 1 (fixed chromosome effects only) explains the data more parsimoniously than Model 2 (fixed chromosome and random block effects), independent of the block size included in Model 2 (Supplemental Fig. 4). When regional effects of varying sizes are simulated, Model 2 provides a substantially better fit to the data, as is the case with our real data. In addition, the best-fitting mixed effects model (i.e., the model with the lowest AIC) is that which includes a block size closest to the true simulated block size (Supplemental Fig. 5).

It should be noted that the mixed model does not explain a large proportion of the variance in substitution rate (~6%) when fitted to data consisting of observations on individual ancestral repeats, as we have presented above. However, it is likely that much of the residual variation is due to the considerable error involved in inferring substitution rates from such small sequences (on average ~200 bp). This is supported by the observation that the proportion of variance explained by the mixed model when fitted to the slightly longer flanking sequences (on average ~362 bp) is higher (~9%). If we assume that there is minimal mutational variation below 50 kb and thus treat all ancestral repeats within a 50-kb window as a single sequence having a single mutation rate, the mixed model, including a term for a 1-Mb regional effect, explains ~25% of the total variation. We consider this to be a reasonable estimate of the proportion of true mutational variation explained by the most parsimonious model in our analyses.

Discussion

Our study provides further evidence for, and clarification of, the regional mutation hypothesis. It appears that the primary scale over which mutation rates vary is subchromosomal and that within-chromosome effects are at least as important as male germ-line effects as a source of mutational variability, although the latter has received substantially more attention in the literature. The evidence for this conclusion is threefold. Firstly, partial autocorrelations suggest that all long-range (>1 Mb) similarity of mutation rates can be explained by "propagation" of similarity of mutation rates across distances of <1 Mb. Secondly, results of the mixed model analysis indicate that within-chromosome mutational variation greatly exceeds variation among chromosomes. Given that chromosomal location of X-linked sequence appears highly conserved between mouse and rat (Gibbs et al. 2004), it is unlikely that the within-chromosome variation we observe could be the result of differences in time spent within the male germ line. Thirdly, comparison of our Models 2 and 3 indicates that the effects of chromosome on mean nucleotide substitution rates are small.

We find little evidence in murids for significant similarity of substitution rates across scales as large as an entire chromosome,

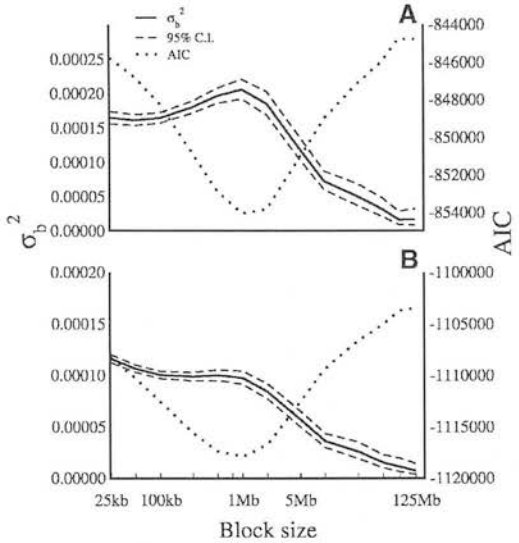

Figure 5. Between-block variation (σ_b^2) in substitution rates within ancestral repeats (*A*) and flanking sequence (*B*). Substitution rates are estimated at all sites. Between-block variances are estimated fitting the chromosome as a fixed effect and the block as a random effect across different block sizes, from 25 kb to 125 Mb. Block sizes are plotted on a $\log_{10}$ scale. The 95% confidence intervals of the between-block variance were as estimated by the lme routine of the nlme package in R. The Akaike Information Criterion (AIC) is shown for each fitted model.

Table 1. Akaike Information Criteria for Model 2 (chromosomal and regional effects) and Model 3 (regional only) when fitted to each of four data sets: nucleotide substitution rates in ancestral repeats and flanking, nonrepetitive sequence, including and excluding the X-chromosome

	All chromosomes		Autosomes only	
	Ancestral repeat	Flank	Ancestral repeat	Flank
Model 2	−854,283.5	−1,118,129	−803,472	−1,052,114
Model 3	−854,133	−1,117,967	−803,349.8	−1,052,005

Both models included a term for a 1-Mb regional effect.

as a previous human–mouse study has indicated (Lercher et al. 2001). A possible explanation is that the mutation pattern has undergone a substantial shift in the lineage leading from the murid common ancestor to human, although how such an event might have occurred remains uncertain. Perhaps a more likely possibility is that the wide divergence between human and mouse simply affords greater power to detect such small effects. Notwithstanding, a recent large-scale study of the synonymous substitution rates at ~15,000 human–mouse gene orthologs supports our conclusion of local similarity extending to 10–15-Mb intervals (Chuang and Li 2004).

It is interesting to note that while our estimates of between-chromosome variation are consistent with previous estimates from murid ancestral repeats (e.g., $\sim 3 \times 10^{-5}$) (Makova et al. 2004), they are lower than the between-chromosome variation estimated at synonymous sites from a recent study (2.7×10^{-4}) (Malcom et al. 2003). However, the average variance of the estimates of mean chromosomal K_S from Malcom et al. (2003) is also somewhat larger than the variance of chromosomal substitution rates we estimate from ancestral repeats (~0.0069 vs. ~0.0025). It seems, therefore, that substitution rates at synonymous sites are considerably more variable than rates within ancestral repeat sequences. This may be a result of selection on some synonymous sites, or interaction between the effects of strong selection on sites adjacent to synonymous sites and context-dependent mutational processes. It is likely, therefore, that the same pattern of variation (within-chromosome mutational variation exceeding variation among chromosomes) would also be evident if rates were estimated at synonymous sites.

Our results raise questions about the biological mechanisms that give rise to new mutations. We suggest that the pattern of variation that we observe could therefore be explained by two, nonmutually exclusive, processes. Firstly, the accuracy of DNA replication may vary regionally along the length of chromosomes. This could elevate or diminish the mutation rate in different regions of the same chromosome. We are, however, unaware of a specific biological mechanism that could produce regionally varying replication accuracy. Secondly, other factors, such as structural alterations and spontaneous degradation of nucleotide bases that are unaffected by DNA replication could contribute substantially to the production of single base-pair mutations. Such alterations could include processes such as the deamination of methylcytosine to thymine or oxidative base damage caused by oxygen free radicals. That the pattern of variation remains the same when considering substitution rates at non-CpG-prone sites (Supplemental Fig. 3) would suggest that CpG-derived mutation is not responsible for much of the regional variation we observe. It is unclear whether those mutations produced by oxidative base damage can be distinguished from mutations derived from other sources, however.

The magnitude of within-chromosomal mutational variation highlights the importance of accounting for regionally varying mutation rates in the identification of putatively functional regions of noncoding DNA. Although the coefficient of regional variation in nucleotide substitution rates we observe is not large (8.75%; 1-Mb regional effects), this still has an impact on the null expectation of conservation of a sequence between two species. As an example, assuming that mouse–rat divergence is normally distributed with a mean of 0.16 and a standard deviation of 0.014, 95% of divergence scores will be in the range 0.132–0.188. The probability of 95% sequence identity of a 100-bp sequence between two species at the lower 95% bound is more than two orders of magnitude larger than the probability of the same sequence at the upper 95% bound. This observation also emphasizes the importance of estimating neutral mutation rates locally. Additionally, our results illustrate that there is likely to be an effect of sampling when estimating average chromosomal substitution rates solely from genic regions. The majority of mammalian genes reside in GC-rich regions (Mouchiroud et al. 1991; Lander et al. 2001); thus even sampling all genes from a chromosome may return a regionally biased estimate of chromosomal evolutionary rate, and any subsamples thereof will potentially exaggerate this bias. Clearly, in order to accurately estimate an average chromosomal mutation rate, one must sample from all regions of a chromosome, not just genic regions, and this could explain some disparities between previous estimates of average X and autosomal substitution rates.

One implication of a subchromosomal mutational scale is that the major process or processes that drive point mutation could be expected to vary across similar scales. One candidate for such a driving process is recombination. Recombination rates have been previously shown to covary with neutral substitution rates in ancestral repeats (Hardison et al. 2003). It is also known that recombination rates in humans are significantly correlated with GC content, probably as a result of biased gene conversion (Kong et al. 2002; Meunier and Duret 2004). Recent results from the highly recombining human pseudoautosomal region provide further evidence that recombination may have an effect on the neutral mutation rate (Filatov 2004). In order to investigate the possibility that recombination rates are related to substitution rates, we collected mouse recombination rate data from a recent comparative study (Jensen-Seaman et al. 2004). These data consist of estimates of local recombination rate in 5-Mb windows across the mouse genome. We estimated average substitution rates for each of these windows from our data. However, we find little evidence for a relationship between mouse recombination rates and mouse–rat divergence; the slope of the regression line of substitution rates on recombination rates is approximately zero (Fig. 6). If recombination is driving mutation in murids, our data suggest that the relationship is not straightforward, on a genome-wide level at least. This conclusion is supported by recent work suggesting that the relationship between recombination rate and nucleotide substitution is at best moderate (Huang et al. 2005). Furthermore, some studies have suggested that the majority of recombinations in humans occur in a comparatively small proportion of the genome (Crawford et al. 2004; McVean et al. 2004). If such recombination "hotspots" also occur in murids, the lack of an observed relationship may be explained, in part, by this effect. For example, if recombination rates vary over scales of kilobases, as opposed to megabases, then any relationship be-

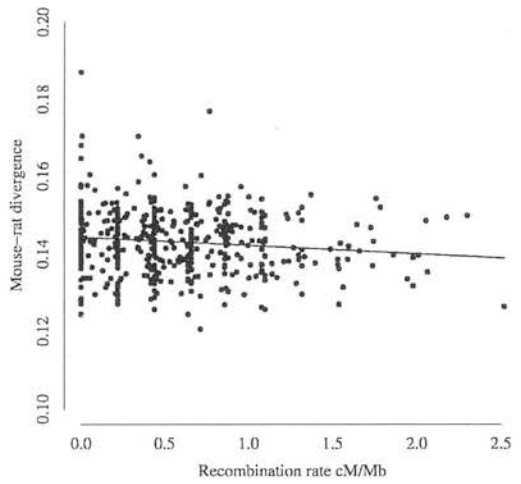

Figure 6. The relationship between mouse–rat divergence and the mouse recombination rate average across 5-Mb windows. The equation of the regression line shown was estimated as $y = 0.144 - 0.002x$.

tween mutation and recombination may be obscured by averaging rates over large genomic distances. In addition, if recombination rates change rapidly over evolutionary time, this may cause problems in deciphering the true nature of any relationship between mutation and recombination, as the latter is measured over much shorter timescales than the former.

One problem to which our data are potentially susceptible is that of gene conversion in repetitive sequence. It has been shown recently that some gene conversion occurs in young *Alu* repeats (Roy et al. 2000). If gene conversion is biased in the direction of the ancestral state, then this will produce a negative correlation between nucleotide divergence and the rate of conversion. The distributions of repeat age within SINEs (results not shown) would suggest that *Alu*/B1 elements differ from the other families of SINEs in that there is a small proportion of *Alu* elements that are younger than other SINE elements. This would suggest either that we have retrieved more *Alu* elements from low-mutating regions or that biased gene conversion toward the ancestral repeat is occurring. If the latter is the case, then there is little we can do to remove this effect from our data, short of locating those elements that are ancestral in a more highly diverged species, for example, human, to minimize the proportion of young *Alu*s in the data set. However, if gene conversion is occurring in some *Alu*s in our data, it appears to have a small effect on our results. The pattern of autocorrelation is practically unchanged if we entirely remove *Alu*s from our data set, as is the ratio of within- to between-chromosome substitutional variation. In addition, previous analyses have concluded that gene conversion in repetitive DNA appears to have small effects on neutral substitution rates at the genomic scale (Makova et al. 2004).

We have shown that the scale of mutational similarity in murids extends from 100 kb to 15 Mb and that the "unit" of mutational variation is no larger than 1 Mb. Our results indicate that, at this scale of regional effect, there exists approximately one order of magnitude more variation in mutation rates within chromosomes than among chromosomes. This has implications for the study of the processes driving mutation and identification of functional noncoding DNA using comparative genomic methods.

Methods

Data

Most mammalian transposable elements can be divided into four broad classes: Short INterspersed Elements (SINEs), Long INterspersed Elements (LINEs), Long Terminal Repeat (LTR) retroposons, and DNA transposons. We identified all SINE, LINE, LTR, and DNA repetitive elements in build 33.1 of the mouse genome using RepeatMasker (http://www.repeatmasker.org/). We identified those repetitive elements that were inserted prior to the mouse–rat divergence as follows. First, 250 bp of sequence upstream and downstream of the identified mouse repeat was extracted. Any repetitive sequence in these flanking sequences was masked, also using RepeatMasker. In order to ensure that matches were achieved using reasonable lengths of sequence, we excluded any element that did not contain at least 50 consecutive bases of unique, nonrepetitive sequence in both its adjacent flanking sequences. Following masking, the remaining unique sequence was compared to the rat chromosome(s) syntenic to the mouse chromosome on which the repeat originated using BLASTN (Altschul et al. 1997). Chromosomal synteny was as defined in Figure 4 of Gibbs et al. (2004). The following criteria were used to accept or reject BLAST hits of pairs of flanking sequence. (1) Hits with *E*-values of $>10^{-5}$ were rejected. (2) Hits were only accepted if both flanks had a single unique match on the same rat contig. (3) To ensure returned BLAST hits were orthologous to the sequence immediately adjacent to the flanks of the original mouse repetitive element, matches of upstream (downstream) flanks were required to extend to within 50 bp of the flank end (start). Fulfilment of these criteria indicated that the sequence surrounding the mouse repeat in question was present in the last common murid ancestor. The region between the outer limits of the matched flanks was then extracted from the appropriate rat chromosome of NCBI build 3.1 of the rat genome and aligned to the original mouse flanks and repetitive element sequence using AVID (Bray et al. 2003). The presence of a clearly orthologous sequence in rat opposite the original mouse repeat in our alignment indicated that the transposable element in question was inserted prior to the mouse–rat divergence.

Estimation of substitution rates

Nucleotide substitution rates were estimated for each ancestral repeat and its flanking sequence, correcting for multiple hits using the Tamura-Nei method (Tamura and Nei 1993). Many transposable elements are GC and CpG rich, and this may affect nucleotide substitution rates, depending on the region of insertion of the element. In addition, analysis of the composition and age of large numbers of repetitive elements in the human genome indicated that element GC content tends to decay over the course of evolutionary time (Lander et al. 2001). This effect violates the assumption of stationarity, common to the majority of models used to estimate substitution rates. It is likely, however, that for moderately diverged species, such as mouse and rat, relatively little GC-content decay will have occurred since the two species split. Of greater concern is the fact that many mammalian repetitive consensus sequences contain hypermutable CpG dinucleotides at a substantially higher frequency than the genome at large. Hypermutability of CpG dinucleotides in vertebrates is well documented and poses a problem for the estimation of substitution rates using ancestral repeats. Following insertion, CpG dinucleotides within elements are by far the most likely sites to mutate. However, ancient elements will have experienced most CpG-related changes prior to mouse–rat divergence, whereas those more recent insertions may appear to be evolving at an

inflated rate because of their comparatively higher CpG content. This effect could produce covariation between the age of insertion and overall divergence, with more CpG-rich recently inserted elements diverging proportionally more than their older counterparts. Although there have been recent advances in incorporating context dependency into models of sequence evolution (Arndt et al. 2003; Siepel and Haussler 2004), in this study we addressed these issues by estimating nucleotide substitution rates in three alternate ways: using all sites, at those sites not preceded by a C or followed by a G (non-CpG-prone sites), and by counting only A↔T and G↔C changes. The latter two categories are likely to be the least affected by CpG context effects and compositional change and allowed us to assess the impact, or otherwise, of these factors on our results.

Mean chromosomal divergence

We calculated the mean chromosomal divergence treating the entire chromosome as a single sequence and summing differences and sites across all elements. Estimates were also corrected for multiple hits using the method of Tamura and Nei (1993). In order to estimate confidence intervals for the average chromosomal substitution rate, we generated 1000 bootstrap data sets for each chromosome. Because adjacent substitution rates are autocorrelated, we bootstrap by 2-Mb blocks to minimize dependence between observations. We calculated the mean chromosomal divergence for each data set, and the bootstrap distribution of these was used to estimate 95% confidence intervals for each mouse chromosome. Bootstrap data sets were generated using the "boot" library in R (R Development Core Team 2004).

Local similarity

To investigate the scale of local similarity of substitution rates, we divided the mouse genome into 5-kb and 100-kb blocks and estimated an average block substitution rate by taking a weighted (by number of sites) average of the substitution rates of all elements found within a block. We then estimated the autocorrelation of substitution rates across blocks. The autocorrelation of substitution rate K in block i with block $i + k$, where k is the order or lag of the autocorrelation, is given by (Box et al. 1994):

$$\rho_k = \frac{\sum_{i=1}^{N-k} (K_i - \overline{K})(K_{i+k} - \overline{K})}{\sum_{i=1}^{N} (K_i - \overline{K})^2} \quad (1)$$

where N is the total number of blocks. In order to provide confidence intervals for the distribution of ρ under the null hypothesis of no relationship between the evolutionary rates of adjacent blocks, we estimated ρ for 1000 data sets in which block order was randomized. Following Matassi et al. (1999) and Lercher et al. (2001), we assessed the impact of local GC content on the observed pattern of autocorrelation using data sets in which blocks were randomized according to their GC content. Because of the nonrandom pattern of insertion of transposable elements, in all cases elements were permuted while maintaining the structure of our original data set, for example, any empty blocks in the real data were maintained as empty blocks in all our randomized data sets. Local GC content was estimated as the average GC content of all masked mouse and rat flanking sequences within a block. Blocks were then assigned to one of several GC-content classes and randomly permuted only with blocks in the same GC-content class, where each GC-content class contained 5% of the data set.

To investigate the mean "unit" of mutational variation, we estimated the partial autocorrelation of substitution rates averaged across 100-kb blocks. Partial autocorrelation between the mean substitution rates in block x_i and block x_{i+k}, where k is the lag, is the amount of correlation that is not explained by the "propagation" of lower-order lags ($k - 1, k - 2, \ldots$). In our case, partial autocorrelation becomes insignificant at the point beyond which all observed similarity of substitution rates can be explained by autocorrelation of rates across smaller distances. All partial autocorrelations were estimated in R. The significance of partial autocorrelations was again assessed using 1000 data sets in which block order was randomized. We estimated partial autocorrelation of substitution rates in both ancestral repeat and flanking sequence up to an interval distance of 5 Mb.

Between- and within-chromosome variation

We estimated a male-to-female mutation rate ratio, α, using the following formula:

$$\alpha = (3R - 4)/(2 - 3R) \quad (2)$$

(Miyata et al. 1987), where $R = X/A$, and X and A are the mean substitution rates at all sites on the X-chromosome and across all the autosomes, respectively.

In order to quantify between- and within-chromosome mutational variation, the data were fitted to a variety of linear models using the nlme library in R (R Development Core Team 2004). Substitution rates in ancestral repeats and flanking sequences were grouped by location into blocks of increasing size from 25 kb to average chromosome size (125 Mb). We then tested the significance of regional effects in explaining variation in the substitution rate by comparing two models:

Model 1

$$y_{ij} = \beta_i + \epsilon_{ij} \quad (3)$$

Model 2

$$y_{ijk} = \beta_i + \beta_i(b_{ij}) + \epsilon_{ijk} \quad (4)$$

In Model 1, the substitution rate y_{ij} is described by an effect of Chromosome i, (β_i), and a random error term (ϵ_{ij}). In Model 2, the substitution rate y_{ijk} is again described by a mean chromosomal rate but also by a mean "regional" rate or effect of block j, b_{ij}, modeled as a normally distributed random effect, nested within the chromosome, that is, as a random variable representing the deviation from the chromosomal mean rate. If substantial regional effects exist, then Model 2 will provide a significantly better fit to the data than Model 1. Both models were fitted to the data using restricted maximum likelihood.

We also tested for significant chromosomal effects by comparing the fit of Model 2 to the data with the following model (Model 3), which includes a term for a random regional effect only:

Model 3

$$y_{ij} = b_i + \epsilon_{ij} \quad (5)$$

If there are significant chromosomal effects, Model 2 will provide a better fit to the data than Model 3. Model 2 and Model 3 were fitted to data both including and excluding the X-chromosome, which is a chromosomal outlier. In this case the data were fitted using "full" maximum likelihood as Model 2 and Model 3 differ

in their fixed effects specification and their log-restricted likelihoods cannot be compared (Pinheiro and Bates 2000).

For all comparisons we used the Akaike Information Criterion (AIC) to assess the fit of the model to the data. The AIC is a model selection criterion that incorporates information about the fit of the model to the data and the model complexity:

$$AIC = -2l(\hat{\theta}|\boldsymbol{y}) + 2n_{par} \quad (6)$$

where $l(\hat{\theta}|\boldsymbol{y})$ is the log-likelihood of the model $\hat{\theta}$, given the data $\boldsymbol{y}$, and n_{par} is the number of parameters in the model (Pinheiro and Bates 2000).

Acknowledgments

We thank Daniel Halligan, Ian White, Toby Johnson, Bill Hill, Gabriel Marais, Alex Kondrashov, and two anonymous referees for helpful comments and discussion. We also thank the Blaxter Lab for the use of their Linux cluster. D.J.G. is funded by a University of Edinburgh postgraduate scholarship.

References

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

Arndt, P.F., Burge, C.B., and Hwa, T. 2003. DNA sequence evolution with neighbor-dependent mutation. *J. Comput. Biol.* **10:** 313–322.

Bailey, J.A., Carrel, L., Chakravarti, A., and Eichler, E.E. 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: The Lyon repeat hypothesis. *Proc. Natl. Acad. Sci.* **97:** 6634–6639.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321–1325.

Box, G.E.P., Jenkins, G.M., and Reinsel, G.C. 1994. *Time series analysis: Forecasting and control*, 3rd ed. Prentice-Hall, Upper Saddle River, NJ.

Bray, N., Dubchak, I., and Pachter, L. 2003. AVID: A global alignment program. *Genome Res.* **13:** 97–102.

Chamary, J.V. and Hurst, L.D. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: Evidence for selectively driven codon usage. *Mol. Biol. Evol.* **21:** 1014–1023.

Chang, B.H.J., Shimmin, L.C., Shyue, S.K., Hewettemmett, D., and Li, W.H. 1994. Weak male-driven molecular evolution in rodents. *Proc. Natl. Acad. Sci.* **91:** 827–831.

Chen, F.C., Vallender, E.J., Wang, H., Tzeng, C.S., and Li, W.H. 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.* **92:** 481–489.

Chin, C., Chuang, J.H., and Li, H. 2005. Genome-wide regulatory complexity in yeast promoters: Separation of functionally conserved and neutral sequence. *Genome Res.* **15:** 205–213.

Chuang, J.H. and Li, H. 2004. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol.* **2:** 253–263.

Crawford, D.C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M.J., Nickerson, D.A., and Stephens, M. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.* **36:** 700–706.

Ebersberger, I., Metzler, D., Schwarz, C., and Paabo, S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70:** 1490–1497.

Ellegren, H., Smith, N.G.C., and Webster, M.T. 2003. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* **13:** 562–568.

Eyre-Walker, A. 1999. Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* **152:** 675–683.

Filatov, D.A. 2004. A gradient of silent substitution rate in the human pseudoautosomal region. *Mol. Biol. Evol.* **21:** 410–417.

Filipski, J. 1988. Why the rate of silent codon substitutions in variable within a vertebrate genome. *J. Theor. Biol.* **134:** 159–164.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.

Huang, S.-W., Friedman, R., Yu, N., Yu, A., and Li, W.-H. 2005. How strong is the mutagenicity of recombination in mammals? *Mol. Biol. Evol.* **22:** 426–431.

Hudson, T.J., Church, D.M., Greenaway, S., Nguyen, H., Cook, A., Steen, R.G., Van Etten, W.J., Castle, A.B., Strivens, M.A., Trickett, P., et al. 2001. A radiation hybrid map of mouse genes. *Nat. Genet.* **29:** 201–205.

Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y.T., Roskin, K.M., Chen, C.F., Thomas, M.A., Haussler, D., and Jacob, H.J. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14:** 528–538.

Keightley, P.D. and Gaffney, D.J. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl. Acad. Sci.* **100:** 13402–13406.

Keightley, P.D., Lercher, M.J., and Eyre-Walker, A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3:** e42.

Khil, P.P., Oliver, B., and Camerini-Otero, R.D. 2005. X for intersection: Retrotransposition both on and off the X chromosome is more frequent. *Trends Genet.* **21:** 3–7.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lercher, M.J., Williams, E.J.B., and Hurst, L.D. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18:** 2032–2039.

Makova, K.D., Yang, S., and Chiaromonte, F. 2004. Insertions and deletions are male biased too: A whole-genome analysis in rodents. *Genome Res.* **14:** 567–573.

Malcom, C.M., Wyckoff, G.J., and Lahn, B.T. 2003. Genic mutation rates in mammals: Local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol. Biol. Evol.* **20:** 1633–1641.

Matassi, G., Sharp, P.M., and Gautier, C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9:** 786–791.

McVean, G.T. and Hurst, L.D. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386:** 388–392.

McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304:** 581–584.

Meunier, J. and Duret, L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21:** 984–990.

Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K., and Yasunaga, T. 1987. Male-driven molecular evolution—A model and nucleotide sequence analysis. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 863–867.

Mouchiroud, D., Donofrio, G., Aissani, B., MacAya, G., Gautier, C., and Bernardi, G. 1991. The distribution of genes in the human genome. *Gene* **100:** 181–187.

Nadeau, J.H. and Taylor, B.A. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci.* **81:** 814–818.

Pinheiro, J.C. and Bates, D.M. 2000. *Mixed-effects models in S and S-PLUS.*. Springer-Verlag, New York.

R Development Core Team. 2004. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 3-900051-07-0.

Roy, A.M., Carroll, M.L., Nguyen, S.V., Salem, A.H., Oldridge, M., Wilkie, A.O.M., Batzer, M.A., and Deininger, P.L. 2000. Potential gene conversion and source genes for recently integrated *Alu* elements. *Genome Res.* **10:** 1485–1495.

Siepel, A. and Haussler, D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21:** 468–488.

Silva, J.C. and Kondrashov, A.S. 2002. Patterns in spontaneous mutation revealed by human–baboon sequence comparison. *Trends Genet.* **18:** 544–547.
Smith, N.G.C., Webster, M.T., and Ellegren, H. 2002. Deterministic mutation rate variation in the human genome. *Genome Res.* **12:** 1350–1356.
Tamura, K. and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10:** 512–526.
Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Webster, M., Smith, N., Lercher, M., and Ellegren, H. 2004. Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol. Biol. Evol.* **21:** 1820–1830.
Willie, E. and Majewski, J. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20:** 534–538.
Wolfe, K.H., Sharp, P.M., and Li, W.H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283–285.

Web site references

http://www.repeatmasker.org/; the program RepeatMasker is available for download from this site.

Received March 2, 2005; accepted in revised form May 3, 2005.