Evaluating the contributions of methylation and transcription to male-biased evolution

Helen Lindsay

A thesis submitted
for the degree of Doctor of Philosophy
of the Australian National University

March 2010

THE AUSTRALIAN NATIONAL UNIVERSITY

John Curtin School of Medical Research
The Australian National University

Statement of Originality

The research presented in this thesis is my own work, which was done in collaboration with others whilst I was enrolled as a PhD student at the John Curtin School of Medical Research at the Australian National University. This work has not been previously submitted for any other degree or award at any other educational institution.

Parts of this thesis refer to published, peer-reviewed journal articles:

1. **Lindsay, H**, Yap, VB, Ying, H and Huttley, GA (2008), 'Pitfalls of the most commonly used models of context dependent substitution.', *Biol Direct* **3**(52)
2. Yap, VB, **Lindsay, H**, Easteal, S and Huttley, G (2010), 'Estimates of the effect of natural selection on protein coding content.', *Mol Biol Evol* **27**(3) 726–734

Helen Lindsay

Abstract

Male biased mutation is thought to be a consequence of mutations introduced during DNA replication. Studies of male bias have generally been restricted to an examination of bias in the total substitution rate rather than the substitution process. In this analysis, the potential contributions of germline sex differences in methylation and transcription to male biased mutation are examined via their effects on the substitution process. It is first shown that one of the post popular methods for modeling the effects of sequence context on nucleotide substitution rates detects an effect of context when none exists, which has important consequences particularly for models that aim to detect natural selection. Transitions involving CpG dinucleotides, which characteristically arise from methylation, are found to make a large contribution to male bias because of CpG frequency differences between the X chromosome and the autosomes. Germline transcription is also found to contribute to male bias, and may completely account for the bias observed in the chimpanzee lineage. These observations indicate that male bias is caused by multiple processes, and the contribution of replication errors is smaller than previously believed.

Acknowledgements

I would like to thank my supervisor Dr Gavin Huttley, for his patience, advice and enthusiasm since my first summer scholarship with him 6 years ago. It is a testament to him that I chose to come back for my honours project, and then my PhD. Our research has always been interesting and enjoyable.

Thank you also to the other members our our research group, past and present, and our collaborators for many productive discussions. In particular, I would like to thank Dr Von Bing Yap for hosting my visit to his lab in Singapore, and for his invaluable suggestions about the experimental design.

Thank you to Ruth Gani. As a result of her generous travel bursary, I was able to attend international conferences where I had many very useful discussions with colleages and formed lasting friendships.

Many thanks to Felix Schill, for his emotional and technical support and limitless patience over the last three years. Thanks also to Drew and Adrian for always being there, and to my friends in the ANU mountaineering club, for helping me do things I'd never imagined doing.

Most importantly, thank you to my parents, who have always encouraged my interest in science, and supported me in my endeavours.

Contents

List of Figures

List of Tables

Chapter 1

Male-biased mutation

Although mutations are the fundamental source of genetic novelty, they are also a source of inherited disease. Normal cellular processes and exogenous toxins pose a constant threat to DNA integrity. Damage that is not repaired, or incorrectly repaired, results in heritable mutations that are passed on when DNA is replicated. Factors that affect the rate of mutation accumulation in the germ line can also affect the probability that a parent transmits a mutation to their offspring that causes disease.

In a variety of taxa including mammals, birds (e.g. Berlin et al., 2006, Ellegren and Fridolfsson, 1997, Kahn and Quinn, 1999, Axelsson et al., 2004), and fish (Ellegren and Fridolfsson, 2003), mutations accumulate more rapidly in the male than the female germ line. This phenomenon is known as male-biased mutation. Weinberg's seminal observation in 1912 that children with achondroplasia tended to be be born later in the sibship than unaffected children led to the discovery of male-biased mutation. Haldane was the first to suggest that if most germ cell mutations arise during DNA replication, males germ cells are likely to have a greater mutation rate than female germ cells due to the greater number of cell divisions they must undergo to reach maturity. Penrose (1957) later determined that the relationship between achondroplasia risk and parental age was predominantly related to the father's age only, supporting Haldane's prediction that replication is an important source of germ line mutation. More recently, paternal age has been linked with a variety of complex disorders, including congenital heart defects (Olshan et al., 1994), schizophrenia (Malaspina et al., 2002), bipolar disorder (Frans et al., 2008) and autism spectrum disorders (Reichenberg et al., 2006).

The male bias in mutation and the mechanisms that cause it have implications for public health, as well as being of major importance to the study of evolution. In the last 50 years, there has been a marked increase in male infertility (Jensen et al., 2002), which may be related to the male-biased mutation rate and the current trend towards delaying parenthood. Determining the etiology of germ line mutations will affect our understanding of fundamental biological questions, including how mutation and the risk of genetic diseases are related to gender and aging, whether age-related fertility declines are related to DNA damage, and how the basic genetic variations upon which evolution acts arise. Potential practical benefits include better models of sequence evolution and more accurate assessment of the recurrence risk for families of children with a genetic disease, as the mechanisms that cause mutations and the developmental stage at which they act affects this probability. The major aim of this project is to assess whether and to what extent two candidate mutagenic mechanisms, methylation and transcription, contribute to the male-biased mutation phenomenon.

In Section 1.0.1, methods for estimating the male mutation bias are discussed, and three factors that potentially cause bias in estimates are considered in Section 1.0.2. The most widely accepted hypothesis to explain male biased mutation is introduced in Section 1.0.3. Experimental results that support and contradict this hypothesis are discussed in Section 1.0.4. Section 1.0.5 reviews sex differences in germ cell biology and how these might contribute to male biased mutation. Finally, Section 1.0.6 explains which mechanisms are examined in this study, and briefly outlines the structure of this thesis.

1.0.1 Estimating the extent of male bias in the mutation rate

Several methods have been used to estimate the male bias in the mutation rate. These include the direct approach of comparing the mutation frequency in mature male and female gametes; the indirect approach of identifying patients with a *de novo* mutation and determining from which parent the mutation was inherited; and the comparative method of considering the evolutionary history of sequences that have spent different periods of time in male and female germ lines. Each approach produces complementary results. The advantages and disadvantages of each method are summarised in Table 1.0.1. Briefly, mutation rate estimates from gametes provide the clearest indication of the germ line mutation spectrum but no indication of whether these mutations are inherited. Studies of patients with de novo mutations provide evidence of the inherited mutation spectrum, which may differ from the germ line spectrum if gametes are subject to natural selection or mutations cause embryonic lethality or infertility, for instance. The comparative approach estimates how mutations have occurred

Table 1.1: Methods for estimating the male bias in the mutation rate.

	Considerations
Germ cells	Provides the most accurate estimate of germ cell mutation. Male germ cells are readily available in large numbers, but there are few mature female germ cells at any time and sampling them is invasive. Mutations in germ cells may not be inherited, e.g. damage can cause infertility or may be repaired in the early embryo. Useful for detecting contemporary sources of mutation.
Pathogenic mutations	Cause a phenotypic change that can influence the probability that the mutation is inherited. Sample sizes are limited by the availability of informative genetic data from patients and their families. Potential ascertainment bias, e.g. if not all mutations come to clinical attention. Possible to detect the parental origin of a single mutation.
Nucleotide substitutions	Possible to consider a large number of substitutions and species. Requires comparison of different loci, consequently estimates are potentially biased by regional variation in mutation pattern or unequal divergence times. Historical sources of mutation may not reflect contemporary sources.

over a long time period, and has the advantage that a large number of mutations from across the genome may be examined.

Initial estimates of the extent of the male mutation bias relied on family studies of X-linked recessive Mendelian diseases. Haldane estimated the male bias in the rate of mutations that cause haemophilia, an X-linked recessive disorder, by estimating the proportion of carrier mothers with affected sons. For an X-linked recessive disorder where most incidences of the disorder are caused by new mutations, the proportion of mothers of affected individuals who are carriers for the mutation is related to the sex-bias in the mutation rate. When the disease-causing mutation occurs in the germ line of the mother, the mother will not be a carrier for the mutation. However, if the disease-causing mutation occurs in the germ line of the maternal grandfather, the mother will inherit the mutation and pass it to all her sons. A method for estimating the sex-bias in the rate of an X-linked recessive disorder based on the frequency of maternal carriers is presented in (Becker et al., 1996).

More recent studies (e.g. Glaser et al., 2000), have estimated the male mutation bias for autosomal dominant Mendelian diseases by comparing the genotypes of affected individuals and their unaffected parents at polymorphic sites linked to the disease-causing mutation. For the parental alleles to be distinguished, the affected individual must be heterozygous for the marker and at least one of the parents homozygous. Sample sizes in these studies are necessarily limited by the number of families of affected individuals where informative genetic data is available.

The comparative method for estimating the male bias in the mutation rate relies on the assumption that the rate of nucleotide substitutions, or fixed changes between lineages, is equal to the mutation rate. According to the neutral theory of molecular evolution, this assumption is true for neutrally evolving sequences, i.e. sequences not subject to natural selection or other forces that alter the fixation probability of a mutation in a predictable way (Kimura, 1983).

Miyata et al. (1987) developed a method for calculating the ratio of the male to female mutation rate (α) by comparing chromosome classes, assuming that the substitution rate of each chromosome reflects the amount of time it spends in the male and female germ lines. The Y chromosome only occurs in males; the autosomes spend equal periods of time in males and females, and the X chromosome spends $2/3$ of its time in females and the remaining $1/3$ in males.

Where the female and male mutation rates are denoted μ_f and μ_m respectively, the substitution rate of the Y chromosome (μ_Y) is μ_m, the substitution rate of the X chromosome (μ_X) is $2/3\mu_f + 1/3\mu_m$ and the substitution rate of the autosomes (μ_A) is $1/2\mu_f + 1/2\mu_m$. The ratio of the male to female substitution rate is denoted $\alpha = \mu_m/\mu_f$. Therefore the ratio of the substitution rate on the X chromosome compared to the autosomes is:

$$\frac{\mu_X}{\mu_A} = \frac{2(2\mu_f + \mu_m)}{3(\mu_f + \mu_m)} = \frac{2(2\mu_f + \alpha\mu_f)}{3(\mu_f + \alpha\mu_f)} = \frac{2(2 + \alpha)}{3(1 + \alpha)} \tag{1.1}$$

Similar calculations yield:

$$\frac{\mu_Y}{\mu_X} = \frac{3\alpha}{2 + \alpha} \quad \text{and} \quad \frac{\mu_Y}{\mu_A} = \frac{2\alpha}{1 + \alpha} \tag{1.2}$$

Analogous results are easily derived for birds, where females are the heterogametic sex.

From 1.1, as α gets infinitely large, the ratio $\frac{\mu_X}{\mu_A}$ of the substitution rates on the X chromosome and the autosomes approaches $2/3$. Early studies using the method of Miyata *et al* found $\frac{\mu_X}{\mu_A}$ was less than the theoretical minimum value of $2/3$. Results from subsequent studies have been variable, and are discussed below. Note that as $\frac{\mu_X}{\mu_A}$ must be positive, the function $f : \frac{\mu_X}{\mu_A} \to \alpha$ maps values of $\frac{\mu_X}{\mu_A}$ from $2/3$ to $4/3$ to values of α ranging from ∞ to 0. That is, small changes in the ratio $\frac{\mu_X}{\mu_A}$ result in large changes in α.

The relationship between the chromosomal substitution rate ratio and α, shown in Figure 1.1, differs for each chromosomal comparison. The three curves in

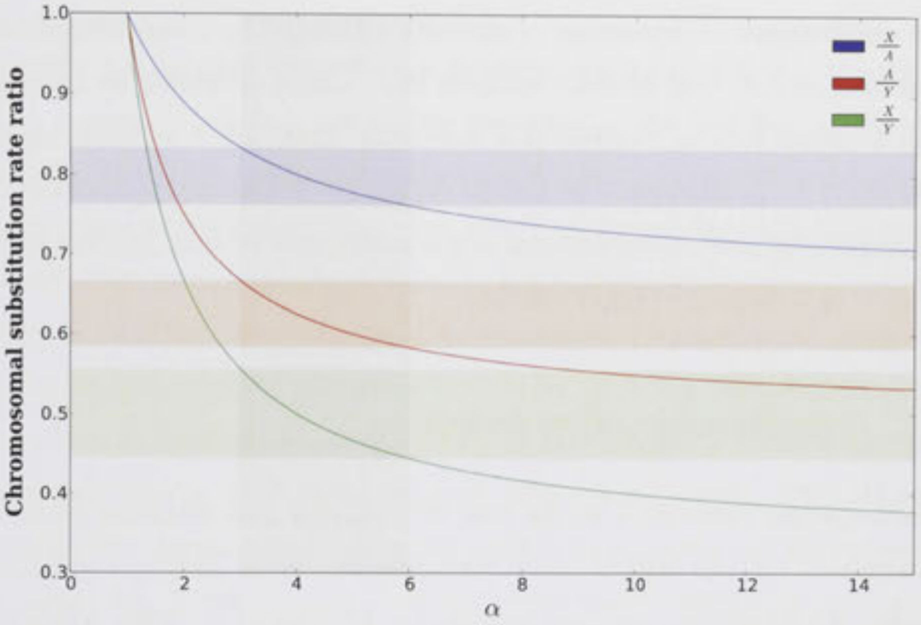

Figure 1.1: Substitution rate ratios and associated α values for different chromosome comparisons. The vertical grey shaded region shows the range from 2.5– 6. α estimates reported for humans have typically been within this range. The horizontal coloured regions show the range of substitution rate ratios that can produce an α estimate in the range 2.5–6. Abbreviations are **X** - X-linked, **Y** - Y-linked and **A** - autosomal.

Figure 1.1 share the property that for small values of α, large fluctations in the chromosomal substitution rate ratio produce small changes in α estimates. This is particularly noticeable for the $\frac{X}{Y}$ comparison, where substitution rate ratios from 0.55– 1 all produce α estimates from $\approx$ 1.5– 4. In contrast, the same range of $\frac{X}{A}$ substitution rate ratios produces α estimates from 1– ∞, with values less than $\frac{2}{3}$ producing an invalid result. The opposite relationship is true of large values of α; very small fluctuations in the chromosomal substitution rate ratio will produce large changes in α. As the curve for $\frac{X}{A}$ comparisons has the smallest gradient at any value of α, α estimates from this comparison can be expected to be the most variable. That is, assuming the substitution rate ratios are equally variable for all chromosome comparisons, α statistics estimated by comparing X-linked and autosomal data will be more variable than values estimated by comparing X- and Y-linked data or Y-linked and autosomal data. All α statistics reported in this study were derived from comparison of X-linked and autosomal data. This comparison was chosen because of the availability of data, and because the variability of α estimates was of interest.

1.0.2 Potential confounding factors

Location-specific differences in the rate of mutation and mutation fixation, if not accounted for, can lead to inaccurate estimates of α. Regional variation in estimates of substitution rate is extensive (e.g. Ellegren et al., 2003, Arndt et al., 2005), but the causes are poorly understood. This poses a particular problem for comparative studies of the male mutation bias, because Miyata's method of α estimation assumes that substitution rate differences between chromosome types are entirely due to the relative period of time spent in the male and female germ line environments. Comparison of an unusually rapidly-evolving X-linked sequence with an unusually slowly-evolving autosomal sequence will lead to underestimation of the "true" male bias, to the extent that this can be defined. Miyata's method also assumes that the loci compared diverged at the same time. This assumption can be violated in closely related species, where differences between the sex chromosomes and autosomes in mutation fixation rates affect divergence times. Factors that potentially confound estimation of α are discussed in the following paragraphs, and their influence on the results presented in this study considered in later chapters.

Regional substitution rate variation

Nucleotide substitution rate varies with the regional nucleotide composition of a DNA sequence, which is commonly measured as the G+C% (e.g. Hurst and Williams, 2000). This variation poses a potential problem for the estimation of α by comparison of the substitution rates of two or more different loci, particularly as dinucleotide frequencies differ between the X chromosome and the autosomes (Huttley et al., 2000). Many studies have attempted to minimise the confounding effect of regional-specific substitution rates by choosing homologous pairs of genes where at least one member of the pair resides on a sex chromosome (e.g. Lawson and Hewitt, 2002, Shimmin et al., 1993, Chang and Li, 1995). However, as translocated pseudogenes adopt the evolutionary pattern of the region they are inserted into (Francino and Ochman, 1999), homologous gene pairs are still affected by regional substitution rate differences. The substitution rates of pairs of homologous introns from the Z and W bird chromosomes are as variable as randomly sampled, non-homologous intron pairs (Berlin et al., 2006).

Natural selection

Natural selection affects the evolution of sex chromosomes and autosomes differently. The mammalian Y chromosome and the avian W chromosome are mostly non-recombining. Selective sweeps on either of these chromosomes will reduce diversity across a large linked region. The efficacy of natural selection can also vary between loci. A recessive mutation on an X or Z chromosome will not be exposed to selection in the homogametic sex, but will be in the heterogametic sex. An autosomal recessive mutation will only be exposed when it is in the homozygous state. Therefore selection can act more effectively on the sex chromosomes. The X chromosome is postulated to have evolved a low mutation rate to avoid exposing mutations in a haploid state in males (McVean and Hurst, 1997). This would have the effect of inflating estimates of male bias. Evidence for a mutation rate reduction on the X chromosome is limited. Whilst the human and chimpanzee X chromosomes show unusually low divergence, this is thought to be a result of either a complex speciation process or a selective sweep (see next section).

Ancestral diversity

The total number of differences between two species can be divided into those differences that occurred after speciation, and those differences that result from the fixation of polymorphic sites that existed at the time of speciation. For closely related species, differences resulting from ancestral polymorphism can constitute a considerable proportion of the total divergence. Burgess and Yang (2008) estimated that 39% of the total divergence between humans and chimpanzees results from ancestral polymorphism. The total divergence d can be expressed as

$$d = 2\mu t + 4N_e\mu \qquad \text{(Kimura, 1983)} \tag{1.3}$$

where μ is the mutation rate, t is the time in years since speciation, N_e is the effective population size of the ancestral population and $4N_e\mu$ is the contribution of ancestral diversity to divergence. The equivalent chromosome-specific values will be denoted with the subscripts X, Y and A for X-linked, Y-linked and autosomal values respectively, for example d_A is the autosomal divergence. From 1.3, it is apparent that the ratio of divergence estimates from different chromosomal classes is affected by differences in speciation time, mutation rate or ancestral effective population size between the classes (see Burgess and Yang, 2008). The X chromosome, Y chromosome and autosomes have different effective population sizes, which poses a potential problem for the estimation of $\alpha = \mu_m/\mu_f$ because different chromosome classes will have different levels of ancestral diversity. Under conditions of neutral evolution where variation in reproductive success is equal for males and females, the effective population sizes N_Y, N_X and N_A of the Y, X and autosomes respectively should reflect the number of copies of each chromosome type segregating in a population at any time, i.e. $N_Y = \frac{1}{4}N_A$ and $N_X = \frac{3}{4}N_A$.

The expected effect of ancestral polymorphism is to decrease the d_Y/d_A and d_Y/d_X ratios and increase the d_X/d_A ratio, leading comparisons of d_Y/d_A and d_Y/d_X to underestimate the male bias and comparisons of d_X/d_A to overestimate the bias. Several studies have compared α estimates from recently diverged species, where ancestral polymorphism has a proportionally greater contribution to divergence, with estimates from more distantly related species (Makova and Li, 2002, Sandstedt and Tucker, 2005, Bartosch-Harlid et al., 2003). The results were consistent with the predicted effects of ancestral polymorphism.

Estimates of α, and the conclusions they support, are sensitive to the method of accommodating the influence of ancestral diversity. For example, Ebersberger et al. (2002) compared α estimates based on Y/X, Y/A and X/A divergence rate ratios for human and chimpanzee sequences. When the ancestral diversity was assumed to equal the present human nucleotide diversity Y/X, Y/A and X/A ratios differed, which contradicts the replication-origin hypothesis. However, if ancestral diversity was assumed to be four times greater than the present human diversity, the Y/X, Y/A and X/A ratios were consistent. The assumption that the ancestral human population size, and thus the ancestral diversity, is much greater than the present diversity is supported by several studies (Burgess and Yang, 2008, Kaessmann et al., 2001). Other studies have tried to correct for the influence of ancestral polymorphism on α estimates by using published speciation time and ancestral population size estimates, (e.g. Nachman and Crowell, 2000), or by attributing all variation in divergence rate ratios to ancestral polymorphism, (e.g. Makova and Li, 2002). Both methods attribute all variation in the divergence rate to one of the three possible sources: the former study attributes all, and the latter study none of the variation, to variation in mutation rate. A more rigorous approach developed by Burgess and Yang (2008) jointly estimates mutation rate, speciation time and ancestral population size.

The relative contributions of ancestral polymorphism and speciation time to the divergence of the human and chimpanzee X chromosomes has been the subject of recent controversy (Patterson et al., 2006, Wakeley, 2008, Burgess and Yang, 2008). The X chromosomes of these species are considerably less diverged than the autosomes, even assuming the effective population size of the X chromosome is $3/4$ that of the autosomes. The initial study by Patterson *et al* attributed the low X chromosome divergence to a complex process of speciation in the ancestral human and chimpanzee lineages, where hybridisation post-speciation led to a low divergence time for the X chromosome. A key factor in support of this hypothesis is that the phylogenetic relationship of human, chimpanzee and gorilla is much less ambiguous on the X chromosome than the autosomes, suggesting a more recent speciation time. Burgess and Yang, however, found that the low X divergence reflects a extremely reduced ancestral X chromosome population size; a possible consequence of a selective sweep. Patterson *et al* estimate α to be 1.9, whilst Burgess and Yang's estimate is 3.0.

1.0.3 The replication origin hypothesis for the male-biased mutation rate

The most widely accepted explanation for the male biased mutation rate is that most mutations arise during DNA replication and consequently more mutations accumulate in the male germ line than the female germ line due to the greater number of replication cycles involved in spermatogenesis than oogenesis. Early during embryogenesis, a population of primordial germ cells is specified. The primordial germ cells migrate to the gonads and proliferate mitotically. In the gonad, female primordial germ cells are known as oogonia, and male primordial germ cells as spermatogonia. Oogonia stop meiotic proliferation and start meiosis before birth (reviewed in Pepling, 2006), whereas spermatogonia temporarily cease mitotic proliferation before birth. From puberty onwards, spermatogonia undergo further, asymmetric mitotic divisions to form one daughter cell committed to differentiating into a cluster of spermatozoa, and one daughter spermatogonium (Zhao and Garbers, 2002). Spermatogonia therefore continue to replicate throughout the reproductive lifetime of adult males; whereas oogenesis involves a fixed number of replication cycles, all but one of which are completed before a female is born.

Mutations can arise at DNA replication through nucleotide misincorporation by the DNA polymerase (copy errors), or through the fixation of premutagenic lesions. For a copy error to persist, it must escape DNA polymerase proofreading and replication associated mismatch repair (MMR). The number of copy errors accumulated per cell division should depend only on the fidelity of the transcription apparatus and post-replicative repair systems. The rate at which lesions are fixed as mutations will depend on the number of lesions encountered, which is a function of the rate of lesion formation and repair, and the method used to process the lesion. The replication-origin hypothesis refers only to copy errors.

If copy errors are the major source of inherited mutation, the expected sex bias in the mutation rate should be proportional to the disparity in the number of replication cycles required to produce mature male and female gametes. This disparity increases with increasing male age, and the replication origin hypothesis predicts that the magnitude of the male bias in mutation rate should increase proportionately. Spermatogonial cells are generally assumed to divide at a roughly constant rate throughout the reproductive lifespan (e.g. Chang et al., 1994, Li et al., 1996). Therefore the male to female ratio of cell divisions, and according to the replication hypothesis also the sex bias in the mutation rate, should increase with age at a constant rate.

On an evolutionary time scale, the male bias in the nucleotide substitution rate predicted by the replication origin hypothesis is influenced by the average age of fathers of offspring that survive and reproduce. This is not necessarily equivalent to the average age at paternity. For example, in several primate species dominant males monopolise dominant females, and dominant females have greater reproductive fitness (Engelhardt et al., 2006). Consequently, the average paternal age is influenced by the age of the dominant males, and the average age of fathers of surviving offspring is even more strongly related.

The disparity in the number of replication cycles undergone by male and female germ cells at the average age of reproduction roughly correlates with the generation time of vertebrate species. Chang et al. (1994) estimated the ratio of male to female germ cell divisions at the average age of reproduction to be 2.0 and 6.2 in rodents and humans respectively. The latter ratio increases to 9.7 if the average human reproductive age is assumed to be 25 instead of 20 years. Similarly, if the average paternal age in humans is estimated to be 15 years, the male to female ratio of cell divisions is expected to be 2.8 (Li et al., 1996). A more detailed analysis of reproduction in humans and wild primate populations estimated the average age at reproduction of humans, chimpanzees and old world monkeys, e.g. rhesus macaque, to be approximately 28, 21 and 12 years respectively (Gage, 1998). Using Gage's estimate of 28 years for the average age of human fathers gives a male to female ratio of cell divisions of 11.8. Accurate estimation of the expected sex-bias in the mutation rate based on the average disparity in male and female germ cell replication cycles requires an accurate understanding of the reproductive biology of the species in question. A species-specific estimate of the frequency of spermatogonial cell division is also necessary, as this can vary even between related species (Parapanov et al., 2007).

Estimates of the expected ratio of male to female replication cycles are predicated on the assumption that a single population of spermatogonial cells complete all of the replication cycles. This is true of rodents, where replication of spermatogonial stem cells is minimised by extensive downstream replication of cells already committed to differentiation (see Ehmcke et al., 2006). Primate testes, however, contain a spermatogonial stem cell population that does not divide under normal circumstances but can be activated following damage to the replicating spermatogonial cells (reviewed in Ehmcke et al., 2006). Recruitment of these normally quiescent cells to the replicating population potentially has a major impact on the sex ratio of germ cell division. The extent to which quiescent primate spermatogonia are activated during the normal reproductive lifespan is unknown, making it difficult to confidently estimate the male bias expected for primates under the replication-origin hypothesis. For several autosomal dominant disorders, the incidence does not increase monotonically with paternal age, but decreases around age 35-40 (Risch et al., 1987). Yoon et al. (2009) suggested this decrease results from the recruitment of quiescent spermatogonia that have not accumulated replication errors into the replicating population to replace damaged cells. Another potential complication for estimating expected values of α arises from evidence of postnatal oocyte replication (Johnson et al., 2004).

1.0.4 Testing the mutation via replication hypothesis

The replication origin hypothesis for the male-biased mutation rate makes clear predictions about the expected magnitude of the male mutation bias and how this should change with age. Specifically, the replication origin hypothesis predicts that the magnitude of the male bias should equal the average ratio of male to female germ cell replication cycles; and the extent of male bias should increase linearly with age, assuming the rate of sperm production does not decrease with age. In reality, these predictions are difficult to test because neither the average ratio of replication cycles that occur in the male and female germ lines, nor how this ratio varies with age is clear. Further predictions of the replication origin hypothesis are that the absolute rate of mutation should depend on the number of replication cycles completed per unit time, and should reflect the fidelity of DNA replication. These predictions are addressed below.

Do mutation rates reflect replication fidelity?

The replication origin hypothesis predicts that mutations rates should reflect the rate of replicative copy errors. A rough argument suggests that replication fidelity is too high to completely account for the human mutation rate if the average age at paternity is assumed to be 20 years. Eukaryotic DNA polymerases make approximately one error per 10^7 nucleotides copied, 0.1% of which are expected to evade repair by the MMR system (Baarends et al., 2001). This is equivalent to less than one mutation on average per diploid human genome per replication cycle. Assuming an average generation time of 20 years, Nachman estimated that approximately 175 mutations have occurred per generation since human and chimpanzee divergence (Nachman and Crowell, 2000). Chang et al. (1994) estimate that by age 20, human spermatozoa have replicated their DNA 205 times and oocytes 33 times. Assuming less than one error occurs per replication cycle and that copy errors are the only source of mutation, an autosome that spends equal time in male and female germ cells should accumulate at most $119 = (205 + 33)/2$ errors per generation, which is considerably less than Nachman's estimate.

Studies of mutation accumulation in transgenic rodents do not support the hypothesis that replication is a major source of mutation. A mutation frequency increase with age has been observed in both dividing and non-dividing rodent tissues (e.g. Ono et al., 2000), indicating that age effects do not necessarily support a replicative origin for mutations. Despite the frequent proliferation of male germ cells, the testis has one of the slowest rates of age-dependent mutation frequency increase (Ono et al., 2000). Indeed, using a different experimental method, Hill et al. (2005) found no age-related mutation frequency change in male germ cells. An age-related mutation spectrum shift towards more transversions has been observed in mice, suggesting that the causes of mutations change with age (Walter et al., 2004). The distribution of mutations also changed with age (Walter et al., 2004), which is not easily explained by the replication origin hypothesis. The mutation spectrum change may be the result of an age-related decline in base excision repair (BER) capacity, which increases the sensitivity of male germ cells to both spontaneous and exogenous mutagens (Cabelof et al., 2002). The importance of BER capacity in determining mutation frequency is suggested by the high mutation frequency in the rodent tissues that also have the lowest BER activity (Cabelof et al., 2002).

The replication origin hypothesis predicts that the number of germ cell replication cycles per year, which is broadly correlated with generation time (Martin and Palumbi, 1993), should determine the substitution rate. However, carnivores evolve at a similar rate to rodents, despite very different generation times (Huttley et al., 2007). Instead, nucleotide substitution rate variation between species suggests a prominent role for endogenous metabolic damage. Multiple regression analysis indicates that metabolic rate is significantly correlated with the rate of synonymous substitution in primates, but generation time is not (Martin and Palumbi, 1993). Further, the sensitivity of nucleotide sites to oxidation is dependent on the surrounding nucleotides, or sequence context, and context-dependent substitution rates vary in accordance with context-dependent sensitivity to oxidation (Stoltzfus, 2008). These results complement the findings discussed above concerning BER activity in rodent germ cells, as oxidative damage is repaired by BER.

The magnitude of the male mutation bias

A male-biased transmission pattern has been observed for several different types of mutation, not all of which are believed to result from replication. Microsatellite mutations (Ellegren, 2000), point mutations (Miyata et al., 1987) and inversions are all male-biased (Becker et al., 1996), which strongly suggests that multiple mechanisms contribute to the overall male mutation bias. The extent of the bias differs between mutation types (Becker et al., 1996) and loci, and is not observed for all types of mutations. Aneuploidies, in particular trisomy 21, are female-biased, as are large deletions (e.g. the deletions leading to neurofibromatosis, Lazaro et al., 1996). The focus of this work is point mutations, as these constitute approximately 70% of known human disease causing mutations (Antonarakis et al., 2000), and the mechanisms that lead to their generation are the least understood.

The magnitude of the paternal mutation transmission bias varies even amongst diseases that result from point mutations. According to the replication origin hypothesis, a paternal mutation transmission bias should be a general feature of monogenic, dominant disorders. However, the proportion of paternally and maternally inherited *de novo* disease causing mutations differs considerably between diseases. This information is summarised in the parent of origin mutation database available at http://www.otago.ac.nz/IGC (Morison et al., 2001). No sex-bias in mutation transmission was found for sporadic point mutations causing Pelizaeus-Merzbacher disease; an X-linked disorder of myelin function

(Mimault et al., 1999), or von Hippel-Lindau disease (Richards et al., 1995); an autosomal dominant condition that predisposes patients to develop certain types of cancer. In contrast, disorders arising from mutations in fibroblast growth factor receptor genes (FGFR1, FGFR2 or FGFR3), including Apert syndrome (Moloney et al., 1996), achondroplasia (Wilkin et al., 1998), Crouzon syndrome (Glaser et al., 2000), Pfeiffer syndrome (Glaser et al., 2000), and Muenke-type craniosynostosis (Rannan-Eliya et al., 2004), show a particularly strong paternal mutation transmission bias. These studies indicate that the disease-causing mutation is exclusively, or almost exclusively inherited paternally. Estimates of the rate of these mutations are orders of magnitude higher than estimates of mutation rate elsewhere in the genome (e.g. Goriely et al., 2003).

Many of the dominant disorders that have a strong paternal transmission bias arise from mutations in a limited number of "hotspots". These hotspots frequently involve CpG dinucleotides, although the recurrent mutations are not always transitions. Two different mutations at a single CpG site, which result in the same amino acid substitution, cause 97% of achondroplasia cases (Wilkin et al., 1998). Mutation hotspots within CpG dinucleotides are also a feature of Apert syndrome (Moloney et al., 1996), neurofibromatosis (Evans et al., 2005), and Rett syndrome (Trappe et al., 2001), although mutations at non-CpG hotspots also contribute to the disease-causing mutation spectra. Several recent analyses convincingly demonstrate that the mutation that causes Apert syndrome also confers a selective advantage on male germ line cells that carry it (Goriely et al., 2003, Qin et al., 2007, Choi et al., 2008). This mutation is believed to cause germ cells that would normally undergo asymmetric cell division to instead undergo symmetric division, thus expanding the germ cell progenitor population that carries the mutation and leading to germinal mosaicism (Qin et al., 2007). Anatomical studies have shown that spermatogonia carrying the Apert syndrome mutations occur in clusters within the testis (Qin et al., 2007, Choi et al., 2008). Younger donors had smaller clusters, suggesting that clonal expansion of mutant cells did not occur prior to puberty (Choi et al., 2008). Similar mosaicism has recently been detected in very old donors for the mutations that cause achondroplasia, suggesting that these mutations might also confer a similar selective advantage on spermatogonia (Giudicelli et al., 2008). Selective mechanisms have been postulated to explain other hotspots of male-biased mutation (Arnheim and Calabrese, 2009).

Estimates of the male mutation bias from comparative studies are extremely variable. Table 1.2 summarises α estimates from several recent comparative studies of mammalian species. The value of α expected for humans under the replication origin hypothesis is so poorly defined that only extremely low or extremely high values of α can refute the replication origin hypothesis. Depending on the average paternal age assumed, the expected α value in humans could be from 2.8 to 11.8. An even greater range is reasonable if ancestral diversity influences estimates. Nevertheless, extreme values of α have been reported. Estimates of human α, prior to correction for ancestral diversity (see Section 1.0.2), range from 1.3 (Ebersberger et al., 2002) to ∞. Even after correcting for ancestral diversity, the lowest estimate of the primate α is 1.8 (Bohossian et al., 2000), lower than predicted by the replication hypothesis. The data used in this study is notable for its unusually extreme CpG depletion (McVean, 2000). Studies of rodents have generally resulted in an α estimate of 2-3, but very high estimates of 8.63–∞ have also been reported (Smith and Hurst, 1999).

Instead of testing for a specific value of α, the concordance between generation time and male mutation bias is often used as an approximate means of testing the replication origin hypothesis. For example, whatever the value of α for humans, it should be greater than that for rodents and birds. Estimates of α are broadly consistent with a generation time effect on the magnitude of male mutation bias, but there are notable exceptions. Male bias in perissodactyls (horses and rhinos) is as strong as in primates (Goetting-Minesky and Makova, 2006), and male bias in birds is almost as high (Hurst and Ellegren, 1998).

Several studies found significant differences between estimates of α from different chromosomal classes (Pink et al., 2009, Smith and Hurst, 1999), supporting the existence of sex-chromosome specific factors that influence mutation. Pink *et al* argue this discrepancy results from a mutagenic effect of recombination in the male germ line. As the Y chromosome is non-recombining for most of its length, mutations arising from recombination in males would increase the rate of nucleotide substitution on the autosomes relative to the Y chromosome. If the magnitude of the male bias arising from copy errors is small the autosomal rate could exceed the Y chromosome rate, as observed in the study by Pink *et al*.

Paternal age effects

If mutations arise at replication, the mutation frequency should increase linearly with age. However, some mutations with a paternally biased transmission pattern do not show the expected age-dependent increase in mutation frequency predicted by the replication origin hypothesis (e.g. Risch et al., 1987, Splendore et al., 2003, Li et al., 2006). Data on X-linked, paternally inherited mutations do not support a paternal age effect (Jung et al., 2003), whilst some autosomal disease-causing mutations show an exponential or even cubic increase in frequency with paternal age (Crow, 1999). One problem with early studies of age-related disease incidence was that the causative mutations were unknown. Crow suggested that as different types of mutation show a different magnitude of male bias, inconsistent age effects resulted from different mutation spectra between diseases (Crow, 2006). More recent studies on known point mutations have also shown that paternal age effects are not always consistent with the predictions of the replication origin hypothesis (e.g. Giudicelli et al., 2008). A striking example is that the observed frequency of the C to G mutation at nucleotide 755 in the FGFR2 receptor in sperm, which causes Apert syndrome, increases with paternal age but the C to A mutation at the same site, which is not known to cause a mutant phenotype, does not (Goriely et al., 2003).

Natural selection is a likely contributor to inconsistent paternal age effects. The incidence in sperm of the mutations that cause achondroplasia increases only modestly with paternal age, and not sufficiently to explain the exponential increase in the incidence of the disorder (Tiemann-Boege et al., 2002, Giudicelli et al., 2008). This may be the result of a selective advantage that gives sperm that carry the mutations greater success at fertilising than non-mutant sperm, or the mutant spermatogonia a proliferative advantage as in Apert syndrome.

A wide range of evidence now indicates that the replication hypothesis does not provide a complete explanation for the male-biased mutation rate. The studies discussed above implicate methylation, age-related changes in mutation spectrum and rate, in particular a decline in BER, chromosome specific effects, recombination and natural selection as factors that contribute to the variability in α estimates. Methodological issues including data choice and alignment method also contribute (Smith and Hurst, 1999). Whether these factors enhance or obscure the male-bias mutation effect is not yet clear. Before the replication origin hypothesis for male-biased mutation is accepted, other potential causes must be excluded.

1.0.5 Sex differences in germ cell biology

Male and female germ cells differ in many respects other than the number of replication cycles each must undergo (see Hedrick, 2007, for review), and some of these differences potentially contribute to the different mutation rates observed in each cell lineage. A male bias in the mutation rate could be the consequence of an event that only occurs, or occurs more frequently in the male germ line. It may also result from some aspect of the male germ cell environment that makes male germ cells relatively more susceptible to mutation than female germ cells. This section will briefly review sex-specific differences in germ cell development. More detailed reviews of the two mechanisms a considered in this work, methylation and transcription, are presented in the relevant chapters.

Gametogenesis involves the same events in males and females, although the timing and stage of development at which they occur differs substantially. Spermatogenesis and oogenesis involve mitotic proliferation, recombination of paternally and maternally inherited chromosomes and the production of a mature haploid gamete during meiosis, followed by quiescence until fertilisation. In females, meiosis begins before birth, and germ cells remain arrested at the meiotic prophase until they are selectively recruited to complete a growth and maturation period around ovulation. The first meiotic division is completed at ovulation, and the second after fertilisation. During the growth period, large amounts of mRNA and protein are produced and stored. The early zygote relies entirely on stored maternal factors, as the zygotic genomic is not activated for transcription immediately (e.g. Picton et al., 1998). In contrast, the production of a mature spermatozoan involves extensive chromatin compaction and removal of non-essential cytosolic factors including the transcription apparatus and repair enzymes. These differences are reflected in the anatomy of the mature gametes: oocytes are one of the largest cells in the body, and spermatozoa one of the smallest.

The mutagenic agents that have been evaluated in the male germ line do not act with equal strength at all developmental stages. Several chemical mutagens act specifically in post-meiotic spermatids and spermatozoa (Allen et al., 1995). Mature sperm have a very compact chromatin structure, with the histones replaced by protamines. When the histone-protamine transition is complete, the DNA is very resiliant to damage (Aitken et al., 2009), but after meiosis and before chromatin compaction is complete the DNA is vulnerable (Marchetti and Wyrobek, 2008). In humans, some histones remain in the mature spermatozoa. The amount of protamine is inversely correlated with measures of DNA damage (Aoki et al., 2006). Human sperm retains more histones than rodent sperm does (in van der Heijden et al., 2008), and is more susceptible to certain types of oxidative damage (Olsen et al., 2005). Differences in vulnerability to damage due to chromatin structure may contribute to between-species differences in the male mutation bias.

Oxidation is one of the major types of DNA damage observed in sperm (Aitken and De Iuliis, 2009). An average of 25 000 oxidative lesions per spermatozoa have been detected in healthy males (Fraga et al., 1991). As the mature sperm lacks repair enzymes, repair of this damage takes place in the zygote and relies on repair enzymes and transcripts produced by the oocyte. The composition of the spermatozoal membrane is particularly vulnerable to oxidative damage, which impairs the capacity for fertilisation (Cocuzza et al., 2007). Thus, oxidative DNA damage in sperm is correlated with decreased fertility and sperm quality. Despite these correlations, strong evidence suggests sperm with oxidative DNA damage are still capable of fertilisation and at least some lesions caused by oxidation escape repair in the zygote. A study of children of whose fathers were smokers and mothers non-smokers found that the paternal smoking prior to conception was associated with a higher rate of childhood cancers, but no decrease in fertility (Ji et al., 1997). Another study found correlations between paternal preconception and pre-natal smoking and childhood cancer, but no corresponding correlations for maternal smoking (Chang, 2009). The effect of smoking was cumulative, suggesting damage affects spermatogonia.

Recombination in male and female germ cells differs in location and rate. In males, recombination tends to be localized towards the telomeres, and towards the centromeres in females (Kong et al., 2002). Though the recombination rate is greater in females, recombination in males appears to have the greater impact on sequence evolution. The human recombination rate is correlated with both G+C% and nucleotide substitution rate, and this correlation is stronger for the male than for the female recombination rate (Webster et al., 2005, Dreszer et al., 2007). This effect is likely mediated by biased gene conversion (BGC); the biased correction of mismatches resulting from recombination in favour of G and C nucleotides over A and T nucleotides. Biased substitutions occur in clusters and tend to be in transcribed regions (Dreszer et al., 2007).

1.0.6 Evaluating the contributions of methylation and transcription to male-biased mutation

Of the sex-biased factors considered above, methylation and transcription were considered in detail in this study. Methylation was examined because CpG sites are a prominent contributor to the male-biased disease causing mutation spectrum, and previous attempts to evaluate the contribution of methylation to the male-biased mutation rate have produced contradictory results (see Chapter 4 on page 77). Transcription was also considered, because male and female gametes differ considerably in their transcription patterns and transcription is known to influence the evolutionary pattern (see Chapter 5 on page 97). Recombination is considered briefly in Chapter 3, and was found to be a potential contributor to observed substitution rate differences but was not evaluated further as the context dependent substitution models used in this study were not well-suited for detecting the effects of recombination.

The causes of male bias were evaluated in this study by estimating the nucleotide substitution rate and process across a genomic-scale data set, and relating this to features of the sequences analysed and to the predictions of the replication origin hypothesis. The following chapter presents the models of sequence evolution used to estimate nucleotide substitution rates. In Chapter 3, regional substitution rate heterogeneity is evaluated with respect to its impact on estimates of male-bias. Analyses of the contributions of methylation and transcription to male-biased evolution follow in Chapters 4 and 5.

The major contributions of this thesis include the discovery of a methodological bias that affects popular models of context-dependent substitution in Chapter 2;

the demonstration in Chapter 5 that male bias estimates from intronic and intergenic primate sequences differ, implying that transcription contributes to male bias and that significant differences in male bias estimates occur even for relatively closely related species; and a comprehensive quantification of the contribution of methylation to male-biased mutation in Chapter 4. It is argued in Chapter 6 that the replication-origin hypothesis for male-biased mutation is too simplistic, and that the α statistic is not ideally suited for estimating male bias.

Table 1.2: Male bias estimates from comparative studies. Abbreviations are **A** = Autosomal, **X** = X-linked, **Y** = Y-linked, **H** = Human, **C** = Chimpanzee, **G** = Gorilla, **B** = Bonobo, **Gi** = Gibbon, **S** = Siamang, **Ms** = Mouse, **R** = Rat, **Hr** = Horse, **FFD** = Fourfold degenerate, **L** = Likelihood, **P** = Parsimony, **D** = Distance, **AA** = amino acid, **AD** = Ancestral Diversity, **Chr** = Chromosome.

	Species	Chr	$\hat{\alpha}$	Data	Method	Notes	Ref
Primates	HC	$\frac{Y}{X}, \frac{Y}{A}, \frac{X}{A}$	**1.3–5.4**	Genomic	P		(Ebersberger et al., 2002)
			2.8–3.2			AD = 4 × present human diversity	
	HCG	$\frac{Y}{X}$	**1.66** (1.19–2.45)	Intergenic	P		(Bohossian et al., 2000)
			1.8 (1.15–2.87)			AD = present chimpanzee diversity	
	H	$\frac{Y}{X}$	**2.1**	Repeats	P		(Lander, 2001)
	HCGBSGi	$\frac{Y}{A}$	**2.23** (1.47–3.84)	Intronic	P	Internal branches, A data from chr 3	(Makova and Li, 2002)
	GB		**4.26**			AD (A) = 19%, AD (Y) = 0	
			5.25 (2.44–∞)			External branches	
	HC	$\frac{X}{A}$	**3.6** (1–∞)	Pseudogenes	L		(Nachman and Crowell, 2000)
		$\frac{Y}{A}$	**0.711–1.95**				
		$\frac{X}{A}$	**8.69–∞**				
		$\frac{Y}{X}$	**16.7–28.8**	*Zfx*/*Zfy*, *Ube1x*/*Ube1y* FFD			
Other mammals	HMsHr	$\frac{Y}{X}$	**3**	Coding	D	Same alpha for each species	(Agulnik et al., 1997)
	Cats	$\frac{Y}{X}$	**4.38** (3.76–5.14)	Zfx/Zfy introns	D		(Pecon Slattery and O'Brien, 1998)
	Mice	$\frac{Y}{A}$	**1.8**				(Geraldes et al., 2008)
		$\frac{X}{A}$	**3.9**				
		$\frac{Y}{X}$	**2.3**				

Chapter 2

Modeling context dependent nucleotide substitution

In this chapter, the context dependent nucleotide substitution models used throughout the following chapters are presented. I address the properties of the models, with emphasis on the behaviour of substitution rate estimates obtained when no sequence context effects exist. The results of this chapter extend the results presented in Lindsay et al. (2008) to consider the context dependent substitution model of Yap et al. (2010).

2.1 Introduction and Theory

The causes of germline mutation can potentially be inferred by considering base substitutions that have accumulated as species have diverged. According to the neutral theory of molecular evolution, when a sequence is not subject to natural selection it accumulates base substitutions at a rate directly proportional to its mutation rate in the germline (Kimura, 1983).

For sequences not subject to natural selection, the nucleotide substitution rate is equal to the germ line mutation rate (Kimura, 1983). Consequently, by relating the estimated substitution process of homologous, neutrally evolving alignments with properties of their constituent sequences such as recombination rate and transcription status, processes that affect the germline mutation rate can potentially be identified.

The substitution rate of single nucleotides is highly affected by their neighbouring nucleotides in mammals (e.g. Hess et al., 1994, Blake et al., 1992, Hwang and Green, 2004) and plants (e.g. Morton et al., 2006). The most well-known example of a context-dependent mutagenic process is the elevated mutation rate of cytosine nucleotides within the context of a CpG dinucleotide. The CpG context affects mutagenesis because cytosine residues within CpG are frequently methylated, and methylated cytosine (mC) nucleotides are highly prone to mutation. In this case, the context-dependence and the high mutation rate occur for different reasons. Other context effects may be a direct consequence of sequence affecting the ability of repair enzymes and other agents to bind to DNA.

The context in which a substitution occurs can provide clues about its origin. Mutagenic agents differ in the types of mutation they characteristically cause and the sequence contexts in which they are likely to cause mutation (Rogozin et al., 2005). Whilst $C \rightarrow T$ substitutions occur in all contexts, those that occur in a CpG context are generally attributed to methylation. Although the context preferences of other mutagenic processes are not as well characterised, differences in the effects of sequence context on nucleotide substitution rates between the sex-chromosome and the autosomes may be useful in resolving the causes of male bias.

Nucleotide substitution patterns have most commonly been estimated either by counting observed changes between homologous sequences, or by modeling sequence evolution as a Markov process. In its simplest form, the parsimonious method ignores the possibility that repeated substitutions have occurred at a single site, and can result in biased estimates of the substitution process even when the sequences being compared are closely related (e.g. Gaffney and Keightley, 2008, Hernandez, Williamson, Zhu and Bustamante, 2007). Counting methods that account for context dependence and multiple substitutions have been developed (e.g. Morton et al., 2009), but are not considered here. Markov models of nucleotide substitution and methods of extending these models to account for the effects of sequence context are considered in the following section.

2.1.1 Markov models of nucleotide substitution

Models of DNA sequence evolution typically assume that evolution proceeds according to a continuous time Markov process. For this assumption to be met, sequence evolution must be *stochastic* (i.e. random as opposed to deterministic) and satisfy the *Markov property*:

Consider a stochastic process and let $X(t)$ be the state the process takes at time t. The process has the Markov property if the transition probabilies satisfy

$$p_{ij}(s,t) := P(X(s+t) = j|X(s_0)\ \forall\ s_0 \leq s) = P(X(s+t) = j|X(s)) \qquad (2.1)$$

where $p_{ij}(s,t)$ is the probability that a transition from state i to state j occurs between times s and t, $\sum_j p_{ij}(s,t) = 1 \ \forall \ 0 \leq s \leq t$ and i in $\{X(t)\}$; and $p_{ij}(t) > 0$ for $t > 0$. Assuming that the probability of a substitution at any instant is very small, the process also satisfies the initial conditions:

$$\begin{aligned} p_{ij}(0) &= 0 \quad \textit{if } i \neq j \\ p_{ij}(0) &= 1 \quad \textit{if } i = j \end{aligned} \tag{2.2}$$

The Markov property implies that evolution is memoryless; that is, the probability of a base substitution only depends on the current state and not the ancestral states. The states of a Markov model of DNA sequence evolution may be, for example, single nucleotides, dinucleotides or triplets of nucleotides, codons, or entire sequences.

DNA sequence evolution is often assumed to be time homogeneous. That is,

$$P(X(t+s) = j|X(t) = i) = P(X(s) = j|X(0) = i) \ \forall s \geq 0 \tag{2.3}$$

The assumption of time homogeneity simplifies the calculation of transition probabilities, as transition probabilities under a time homogeneous model depend only on the elapsed time, rather than the start time as well as the elapsed time. Several methods of relaxing the homogeneity assumption whilst retaining its computational advantages have been developed. These include a method that allows each branch of a phylogenetic tree to evolve according to a separate time-homogeneous process, (Galtier and Gouy, 1998) and a method that allows sequences to switch between different substitution processes over time according to a hidden Markov model (Whelan, 2008).

Consider a homogeneous Markov process. The probability $p_{ij}(t)$ can be expressed in terms of conditional probabilities:

$$\begin{aligned} p_{ij}(t+s) &= \sum_k P(X(t+s)=j, X(t)=k|X(0)=i) \\ &= \sum_k P(X(t)=k|X(0)=i)P(X(t+s)=j|X(t)=k) \\ &= p_{ik}(t)p_{kj}(s) \end{aligned} \tag{2.4}$$

(Chapman-Kolmogorov equation)

The Chapman-Kolmogorov equation can be used to derive an expression for the substitution probabilities as a function of time. First, let $P(t)$ be the matrix of $p_{i,j}(t)$ terms and define the *instantaneous rate matrix*, Q as

$$\begin{aligned} Q &= P'(0) \\ \textit{i.e. } q_{ij} &= \lim_{\Delta t\to 0} \frac{p_{ij}(\Delta t) - p_{ij}(0)}{\Delta t} \end{aligned} \tag{2.5}$$

To ensure that probability is conserved, the rows of Q must sum to zero, which can be achieved by defining

$$q_{ii} = -\sum_{i,i\neq j} q_{ij} \tag{2.6}$$

Substituting the initial values (2.2) for P(0) into 2.5 gives

$$q_{ii} = \lim_{\Delta t\to 0} \frac{p_{ii}(\Delta t) - 1}{\Delta t} \quad \textit{and} \quad q_{\substack{ij,\\ i\neq j}} = \lim_{\Delta t\to 0} \frac{p_{ij}(\Delta t)}{\Delta t} \tag{2.7}$$

The Q matrix describes the propensity for substitutions to occur. Entries of a Q matrix will be described as instantaneous, relative substitution rates, or simply as a substitution rate when the substituted motifs are specified.

Using 2.4, and noting that because of the initial condition, $P(0) = I$, where I is the identity matrix,

$$\begin{aligned} P_{ij}(t+\Delta t) - P_{ij}(t) &= \sum_k P_{ik}(t)P_{kj}(\Delta t) - P_{ij}(t) \\ &= \sum_{k\neq j} P_{ik}(t)P_{kj}(\Delta t) - P_{ij}(t) + P_{ij}(t)P_{jj}(\Delta t) \\ &= \sum_{k\neq j} P_{ik}(t)P_{kj}(\Delta t) - P_{ij}(t)(I - P_{jj}(\Delta t)) \end{aligned}$$

Dividing both sides by Δt and taking the limit as $t \to \infty$

$$P'_{ij}(t) = \lim_{\Delta t\to 0}\left[\frac{\sum_{k\neq j} P_{ik}(t)P_{kj}(\Delta t) - P_{ij}(t)(I - P_{jj}(\Delta t))}{\Delta t}\right]$$

The limit and sum may be interchanged because the sum is finite. Interchanging these and using the identities derived in 2.7,

$$\begin{aligned} P'_{ij}(t) &= \sum_{k\neq j} P_{ik}(t)q_{kj} + P_{ij}(t)q_{jj} \\ &= \sum_k P_{ik}(t)q_{kj} \end{aligned} \tag{2.8}$$

In matrix form, 2.8 is written $P'(t) = P(t)Q$

A similar construction to the above gives $P'(t) = QP(t)$. Using the initial conditions, these equations can be solved to give

$$P(t) = e^{Qt} \tag{2.9}$$

Two additional simplifying assumptions are frequently made: *stationarity*, which means that the motif frequencies $\{\pi_i\}$ do not change over time, and *reversibility*, which means that the frequency of i to j substitutions is equal to the frequency of j to i substitutions, i.e.:

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \tag{2.10}$$

For a substitution model applied to single nucleotides, reversibility can be achieved by expressing the q_{ij} in terms of a symmetric rate component, $r_{i,j} = r_{j,i}$, and a motif frequency component, π_j

$$q_{ij} = r_{i,j}\pi_j \tag{2.11}$$

It will be shown in Section 2.1.5 that the same condition can be applied to ensure that the context-dependent models used in this study are reversible.

The assumption of reversibility guarantees that the matrix exponential (2.9) can be computed by spectral decomposition (see Schranz et al., 2008), and simplifies the computation of the likelihood of a substitution model given a sequence alignment (Felsenstein, 1981, and see Section 2.1.3).

Fitting a model of sequence evolution to an alignment of DNA sequences related by a phylogenetic tree involves specifying the form of the Q matrix, and finding the combination of parameter values that would make observing the alignment most probable, assuming the model is correct. The assumption of stationarity means that the motif frequencies at the root of the tree do not need to be optimised in the substitution model, as they are equal to the motif frequencies in the observed sequences.

2.1.2 The General Time Reversible model

The General Time Reversible (**GTR**) nucleotide substitution model, defined below, is the most general stationary, homogeneous and reversible model of single nucleotide evolution, and forms the basis for the context-dependent models used in this study. Each type of nucleotide exchange is allowed to occur at a different rate. The columns and rows in the **GTR** model below are ordered [A,C,G,T]. Diagonal entries are calculated as defined above.

$$Q = \begin{bmatrix} - & r_{A,C}\,\pi_C & r_{A,G}\,\pi_G & r_{A,T}\,\pi_T \\ r_{A,C}\,\pi_A & - & r_{C,G}\,\pi_G & r_{C,T}\,\pi_T \\ r_{A,G}\,\pi_A & r_{C,G}\,\pi_C & - & r_{G,T}\,\pi_T \\ r_{A,T}\,\pi_A & r_{C,T}\,\pi_C & r_{G,T}\,\pi_G & - \end{bmatrix}$$

2.1.3 Estimating the parameters of a model of evolution

Parameter estimates are generally calculated according to the criterion of maximum likelihood. If each site of a DNA sequence is assumed to evolve independently, the likelihood of the sequence can be calculated by summing the likelihood of each site. Felsenstein (1981) developed a "pruning" algorithm for calculating the likelihood of a sequence alignment for a given model and phylogenetic tree. For a reversible model of nucleotide substitution with independent branch lengths, the likelihood is independent of the placement of the root (proof in Felsenstein, 1981).

2.1.4 Context dependent models of evolution

The first context dependent models were developed to model the evolution of codons (Goldman and Yang, 1994, Muse and Gaut, 1994) and sequences that form stem-loop structures in RNA (Schoniger and von Haeseler, 1994). These models, and their parameterisations, have influenced subsequent model development. The codon alphabet naturally segregates nucleotides into non-overlapping triplets of nucleotides, and nucleotide substitution models can easily be extended to model substitutions between independent codons. The specification of motif frequencies in a codon model is less straightforward than in a nucleotide substitution model (see Yap et al., 2010, Lindsay et al., 2008). The original codon models differ in terms of their specification of the equilibium state probabilities. The implications of these differences were examined in (Lindsay et al., 2008) and are briefly revisited in Section 2.3.

Codon models have been extended to consider neutral, context-dependent effects that occur within and across codon boundaries (e.g. Jensen and Pedersen, 2000, Pedersen and Jensen, 2001, Saunders and Green, 2007). Other context-dependent models have been developed to account for the effects of neighbouring nucleotides on substitution rates in non-coding regions (Hwang and Green, 2004). When the influences of both the left and right neighbouring sites are considered, the probability of a substitution depends on the current state of its neighbours, and analytical calculation of the likelihood becomes intractable. Instead, parameter values must be estimated using a method such as Markov Chain Monte Carlo (MCMC). This dramatically increases the computational effort needed to estimate parameter values. The most general context-dependent models have been applied to only small numbers of alignments (Hwang and

Green, 2004), and some are only applicable to pairs of sequences (Jensen and Pedersen, 2000). Simpler context-dependent models consider context effects on one side only (Siepel and Haussler, 2004, Saunders and Green, 2007). The likelihood function can be expressed in terms of conditional probabilities when the dependence occurs in one direction only (Baele et al., 2008), and analytic parameter estimation is still possible.

Arndt *et al* have developed a series of context-dependent substitution models where parameters are estimated using an approximation to the likelihood (see Arndt, Burge and Hwa, 2003). In these models, substitutions occur according to two processes: one context-dependent and one context-independent, which act simultaneously on a sequence. The context-dependent process allows the rate of substitution of a nucleotide to depend on the identity of its two flanking neighbours. An approximation for the trinucleotide frequencies is used to make likelihood estimation tractable. Arndt, Burge and Hwa's (2003) original model allows the estimation of stationary substitution frequencies for a given transition rate matrix, or transition rates given stationary nucleotide and dinucleotide frequencies. This model was extended by Arndt, Petrov and Hwa (2003) to the case of a star phylogeny with a known ancestral sequence, allowing transition rates to be estimated without assuming the observed sequences are at compositional equilibrium. A further generalisation of this model allows the ancestral nucleotide frequencies to be estimated by maximum likelihood for a star phylogeny of three or more sequences (Arndt, 2006). However, the approximate likelihood technique remains computationally intensive (Baele et al., 2008). The most recent extension, by Duret and Arndt (2008), is non-stationary and non-reversible, but likelihoods were estimated by MCMC.

Limited attention has been given to characterising how Q matrices estimated with different models vary. Siepel and Haussler (2004) have compared parameter estimates obtained using reversible and non-reversible models, and by considering alignments as composed of independent or overlapping N-tuples. The parameter estimates produced using the independent sites model were described as "close to optimal" when compared with the estimates produced using the overlapping sites model. The Q matrices estimated using reversible and non-reversible models were similar, except for CpG transversions. This discrepancy may be a consequence of Siepel and Haussler's (2004) treatment of the motif frequencies (see Section 2.3.1). That the other estimates were similar suggests that good approximations of the substitution rates can be obtained with a reversible substitution model, at least for relatively shallow evolutionary time-depths. A further reason for considering reversible substitution models is that some of the algorithms used for matrix exponentiation are vulnerable to error when the matrix being exponentiated represents a non-reversible substitution process, and these vulnerabilities have only recently been characterised (Schranz et al., 2008).

2.1.5 The CpG baseline model

The context-dependent substitution models that are used in this study differ from the original codon substitution models (Lewontin, 1989, Goldman and Yang, 1994, Muse and Gaut, 1994) in their definition of the motif frequencies. Nucleotide substitution was modeled as a stationary, homogeneous and reversible Markov process acting on dinucleotide motifs. Aligned columns of dinucleotides were considered independent, allowing the likelihood to be calculated using Felsenstein's (1981) algorithm. Simultaneous double nucleotide substitutions were disallowed. The importance of simultaneous double nucleotide substitutions to the substitution rate in primates is controversial (e.g. Smith et al., 2003, Whelan and Goldman, 2004), and needs revisiting using the new conditional nucleotide frequency (CNF) model form presented in (Yap et al., 2010) and considered for the dinucleotide case in Section 2.3.1. However, such an investigation should also consider the potentially confounding affect of the complex speciation of the human and chimpanzee lineages (see discussion in Chapter 5). In the models used in this study, a simultaneous dinucleotide substitution would occur via two single-nucleotide steps. If double nucleotide substitutions were prevalent, their effect on the model parameters would be to increase the estimated rate of each of the single nucleotide steps.

Three different methods of specifying the motif frequency component of a Q matrix parameter are compared in the Results section. As the CNF specification (Yap et al., 2010) was used throughout the following chapters, the context-dependent model is defined below with respect to this form. Of the models considered, the interpretation of context-dependent rate estimates is most straightforward using the CNF form, as the effects of context-dependent substitution rates and context-dependent motif frequencies are confounded by the other models (see Section 2.3.1).

In a CNF model, Q matrix entries consist of rate parameters. e.g. $r_{i,j}$, multiplied by the conditional probability of the nucleotide resulting from the substitution, given the identity of its neighbouring nucleotide. The conditional probability of a C nucleotide given that the 3′ neighbouring nucleotide is a G will be written as $\pi_{C|-G}$. The **GTR** parameters were included in every substitution model considered, and context dependent terms introduced as scaling factors adjusting the value of the relevant **GTR** parameter.

The Q matrix for a **GTR** dinucleotide substitution model with an additional parameter for substitutions between CpG and TpG is defined below. This parameterisation with the addition of a parameter for substitutions between CpG and CpA dinucleotides is frequently used in later chapters, and will be referred to as the CpG baseline model.

Let $q_{i_1i_2,\, j_1j_2}$ be the Q matrix parameter representing a substitution from dinucleotide i_1i_2 to dinucleotide j_1j_2, where $i_1, i_2, j_1, j_2 \in \{A, C, T, G\}$. The dinucleotide substitution model is described supposing that the substituted nucleotide occurs at the first position. Substitutions that occur at the second position are defined similarly.

$$
q_{i_1i_2,\, j_1j_2} = \begin{cases}
0 & i_1 \neq j_1 \text{ and } i_2 \neq j_2 \\
\pi_{j_1|-j_2} & i_1 \text{ and } j_1 \text{ differ by } T \leftrightarrow A \\
r_{A\leftrightarrow C}\, \pi_{j_1|-j_2} & i_1 \text{ and } j_1 \text{ differ by } A \leftrightarrow C \\
r_{A\leftrightarrow T}\, \pi_{j_1|-j_2} & i_1 \text{ and } j_1 \text{ differ by } A \leftrightarrow T \\
r_{C\leftrightarrow G}\, \pi_{j_1|-j_2} & i_1 \text{ and } j_1 \text{ differ by } C \leftrightarrow G \\
r_{C\leftrightarrow T}\, \pi_{j_1|-j_2} & i_1 \text{ and } j_1 \text{ differ by } C \leftrightarrow T \text{ and} \\
 & \{i_1i_2, j_1j_2\} \neq \{\text{CpG ,TpG}\} \\
r_{C\leftrightarrow T}\; r_{CG\leftrightarrow TG}\, \pi_{j_1|-j_2} & \{i_1i_2, j_1j_2\} = \{\text{CpG, TpG}\}
\end{cases}
$$

To see that the CNF definition of motif frequencies results in a reversible substitution model and a symmetric matrix of instantaneous rates, note that

$$\pi_{j_1|-j_2} = \frac{\pi_{j_1 j_2}}{\pi_{-j_2}}$$

Where π_{-j_2} is the probability that the $3'$ neighbour of a substituted nucleotide is j_2. Assuming as above that the substitution occurs at the first position, $\pi_{-j_2} = \pi_{-i_2}$ as $i_2 = j_2$.

The reversibility condition for a dinucleotide model, $\pi_{i_1 i_2} Q_{i_1 i_2, j_1 j_2} = \pi_{j_1 j_2} Q_{j_1 j_2, i_1 i_2}$ implies

$$\begin{aligned}
\pi_{i_1 i_2}\, r_{i_1 i_2, j_1 j_2}\, \pi_{j_1|-j_2} &= \pi_{j_1 j_2}\, r_{j_1 j_2, i_1 i_2}\, \pi_{i_1|-i_2} \\
\text{i.e. } \frac{r_{i_1 i_2, j_1 j_2}\, \pi_{j_1|-j_2}}{\pi_{j_1 j_2}} &= \frac{r_{j_1 j_2, i_1 i_2}\, \pi_{i_1|-i_2}}{\pi_{i_1 i_2}} \\
\frac{r_{i_1 i_2, j_1 j_2}}{\pi_{-j_2}} &= \frac{r_{j_1 j_2, i_1 i_2}}{\pi_{-i_2}} \\
\implies r_{i_1 i_2, j_1 j_2} &= r_{j_1 j_2, i_1 i_2}
\end{aligned}$$

Two other widely-used motif probability specifications are considered in this chapter. The nucleotide frequency (NF) specification is similar to the CNF form, except that the motif frequency component of a Q matrix parameter is the frequency of the *nucleotide* resulting from the substitution, independent of its context. In the tuple frequency (TF) specification, the motifs are the *dinucleotides* that results from the substitution. For example, $r_{CG \leftrightarrow TG}$ would be weighted by $\pi_{T|-G}$ in a CNF model, by π_T in an NF model, and by π_{TG} in a TF model.

2.1.6 Interpretation and comparison of parameter estimates

Estimates of substitution rate and time are confounded; a process where substitutions occur at rate r for time t is indistinguishable from a process where substitutions occur at rate $r/2$ for time $2t$ because substitution probabilities depend only on the product of the rate and time (see Equation (2.9)). It is redundant to estimate values for a set of parameters that span Q, such as the **GTR** parameters, as well as the branch lengths. Consequently only five of the six **GTR** parameters are included in the CpG baseline model definition, and the remaining parameter value is set to one. Following parameter estimation, the Q matrix is conventionally scaled so that the expected number of substitutions per site is one, i.e. $-\sum_i \pi_i q_{ii} = 1$, allowing the branch lengths to be interpreted as the expected number of substitutions per site.

Scaling the Q matrix in terms of motif frequencies can introduce correlations between sequence composition and relative rate estimates. Consider a sequence with a highly elevated rate of CpG substitution and a very high frequency of CpG dinucleotides. In this sequence, most substitutions occur at CpG dinucleotides. Scaling the Q matrix so that the average expected number of substitutions per site equals one means that the the relative rate estimate for CpG substitutions must be approximately one.

Determining whether differences between the X chromosome and the autosomes are caused by intrinsic differences in substitution propensities required that the Q matrix be scaled independently of the motif frequencies so that Q matrix entries for different alignments were directly comparable. The intuitive approach used in this study was simply to remove the π terms from the scaling condition, so that the entries of a Q matrix sum to one.

Comparing substitution patterns

Four general approaches have been used in previous studies for quantifying substitution pattern variation:

- **Nested likelihood ratio (LR) tests** assess substitution pattern heterogeneity in terms of its ability to improve model fit.
- **Matrix distance methods** estimate the difference between two or more transition matrices. Existing methods comparing transition matrices do not identify which types of changes exhibit rate heterogeneity, although Weiss and von Haeseler (2003) note that the approach could be extended to do so.
- **Comparison of substitution frequencies** Observed and expected substitution counts have been compared using χ^2 (e.g. Zheng et al., 2007) and G-tests (e.g. Pacholczyk and Kimmel, 2005). Substitution frequencies can also be derived from a Markov transition matrix, and used to estimate the contribution of a given type of substitution to the total divergence (e.g. Goldman and Yang, 1994, Singh et al., 2006, Arndt and Hwa, 2005).
- **Pairwise comparison of instantaneous transition rates** can be performed using similar statistical methods as for the comparison of substitution frequencies.

In this study, several methods of evaluating variation in the substitution process are used. The relative importance of rate parameters for improving model fit is assessed using LR tests. Instantaneous rate parameters are directly compared using Wilcoxon signed-rank tests. Mean values for X-linked and autosomal parameter estimates are also compared indirectly via the α statistic. Confidence intervals for α estimates are derived by mapping bootstrap confidence intervals for substitution rate ratios to α statistics. To disentangle the influences of motif frequencies and nucleotide mutability on branch length estimates, instantaneous substitution rates and substitution frequencies are compared, where a substitution frequency is the total contribution of a given substitution type towards the branch length estimate, defined as in Goldman and Yang (1994).

The effect of motif specification on the ability of dinucleotide models (Lindsay et al., 2008) and codon models (Yap et al., 2010) to correctly model a substitution process with no context effects have been reported previously. The results below update the dinucleotide analysis to include the CNF model.

2.2 Methods

2.2.1 Data

Single-copy sequences from large syntenic regions were chosen for analysis in this project, with the aim of minimising the impact of alignment errors and ectopic recombination on parameter estimates. The confounding influence of natural selection was minimised, as much as possible, by considering only non-coding sequences, and using a large data set. Two data sets of alignments of human, chimpanzee and macaque sequences were used throughout these analyses - the "intronic" and the "intergenic" data set. A subset of these data sets, the "flanking pair" data set, is used in Chapter 5. Any additional data are described in the relevant chapter.

Alignments of human, chimpanzee and macaque one-to-one orthologous sequences were sampled from Ensembl version 49. An intronic data set of 3319 alignments and an intergenic data set of 7662 alignments were selected. The intronic data set was created by selecting all available alignments to human protein-coding gene sequences from chromosomes (as opposed to unanchored "scaffolds"). Annotated exons and CpG islands, as well as simple (di- or trinucleotide) repeats and alignment columns containing a gap character were removed. The original dinucleotide reading frame was preserved so that artificial dinucleotides were not created. Where the original alignment was from the minus (i.e. non-coding, template or transcribed) strand, the reverse complement of the original alignment was created, so that all final alignments were of the non-transcribed strand. Sequences from consecutive introns from the same original gene alignment were concatenated, again preserving the original dinucleotide frame, and divided into blocks of length 50 000 aligned nucleotides. Some genes were sufficiently large that several non-overlapping alignments of 50 000 nucleotides could be made from the same gene. In these cases, all of the alignments were included in the final intronic data set.

The intergenic alignment was created by excluding all annotated human genes from the whole genome alignment of human, chimpanzee and macaque sequences. Regions where the alignment was ambiguous, i.e. where more than one alignment was available, were excluded. Simple repeats, CpG islands and alignment gaps were removed as for the intronic data set, and the remaining alignments were divided into non-overlapping blocks of 50 000 aligned nucleotides.

2.2.2 Model fitting

Details of the model parameterisation were discussed in Section 2.1.5. A standard, single nucleotide GTR substitution model was used in Section 2.3.1 to calculate parameter estimates that were then used to simulate alignments. In all other cases, including models with no dinucleotide relative rate parameters, the states of the model were dinucleotide motifs. Models were fitted to alignments

Table 2.1: Notation

Abbreviation	Description
$i:j$	A complementary nucleotide pair consisting of nucleotide i paired with nucleotide j
GTR	The General Time Reversible nucleotide substitution model (Section 2.1.2)
$r_{i\leftrightarrow j}$ / $r_{i,j}$	the instantaneous rate of substitution between motifs i and j in a reversible Q matrix. These terms are used interchangeably, and the same notation is used for dinucleotide parameters.
$\hat{r}_{i\leftrightarrow j}$	The maximum likelihood estimate of the parameter described above. The ˆ notation is also used for estimates of the α statistic for male bias
$\pi_{X\|-Y}$	The probability of nucleotide X, given that the 3′ neighbour of X is a Y nucleotide
$AB \leftrightarrow AC$	A substitution of AC for AB, or vice versa, where A,B and C are arbitrary nucleotides
$AB \leftrightarrow NN$	A substitution of AB for any motif differing by an instantaneous change, or vice versa, where A,B are arbitrary nucleotides and N is any nucleotide
$r_{AB\leftrightarrow AC}{}^{*}$	A pair of strand complementary substitution parameters, e.g. $r_{TG\leftrightarrow CG}$ and $r_{CG\leftrightarrow TG}$
CpG baseline	Q matrix consisting of the 6 parameters of the GTR model and parameters, and parameters for instantaneous CpG transitions, i.e. $TG \leftrightarrow CG$ and $CA \leftrightarrow CG$. The CpG baseline model is applied to an alphabet of dinucleotide motifs
$GTR + dinuc$	A dinucleotide substitution model containing the GTR parameters and one of the 48 parameters representing an instantaneous substitution between dinucleotides. The '+' notation is used to indicate the inclusion of an additional parameter in a substitution model.

of human, chimpanzee and macaque sequences, with an unrooted tree topology. Modeling notation is summarised in Table 2.1.

Model fitting was performed using the PyCogent software toolkit, version 1.4dev (Knight et al., 2007). Likelihood was used to evaluate model fit, and was computed using the standard algorithm (Felsenstein, 1981). The optimal parameter estimates reported are those that maximised the likelihood of a substitution model for a given alignment. Where not otherwise specified, parameter estimates were calculated by fitting a CpG baseline model to the intergenic data set. Optimisation was performed using the built-in PyCogent simulated annealing and Powell optimisers. A global optimisation was initially performed using simulated annealing, with tolerance set to 1, followed by a local optimisation using Powell. A maximum of 100 000 likelihood evaluations were allowed, and optimisation was restarted up to 10 times. Rate parameter values were bounded from below by zero and from above by 100. Three alignments from the intergenic data set were consistently poorly optimised and were excluded from consideration throughout this work.

2.2.3 Simulation of alignments to evaluate the effects of different motif frequency specification

One alignment with a low G+C% (ENSG00000118946 from chromosome 13, $G + C = 32\%$), one alignment with an average G+C% (ENSG00000115423 from chromosome 2, $G + C = 39\%$) and one alignment with a high G+C% (ENSG00000175866 from chromosome 17, $G + C = 59\%$) were selected from the intronic data set.

To evaluate the influence of motif frequency specification on the apparent influence of sequence context, alignments were simulated that matched the dinucleotide composition of the real alignments, but evolved according to a single nucleotide process. For each of the original alignments, a standard, single nucleotide GTR substitution model was fitted, and then a dinucleotide GTR substitution model with CNF motif frequencies and parameter values fixed as the maximum likelihood estimates from the single nucleotide substitution model was used to simulate 1000 new alignments each of length 10 000 aligned nucleotides. $GTR + CG \leftrightarrow NN$ models were fitted to the simulated alignments, where $CG \leftrightarrow NN$ includes all instantaneous (single-nucleotide) substitutions involving a CpG dinucleotide. Motif frequencies were specified as the frequency of either the substituted dinucleotide (TF), the substituted nucleotide (NF) or the substituted nucleotide conditional on the unchanged nucleotide (CNF).

2.3 Results

2.3.1 Effect of motif specification on estimated effects of context

Context-dependent substitution models aim to measure the extent to which the substitution rate of a nucleotide depends on the identity of other nucleotides. If the nucleotide substitution rate is unaffected by other nucleotides, no context effects should be detected and the estimated rates of single nucleotide substitutions should be equal to rates estimated with an independent sites model, i.e. the context-dependent Q matrix parameters should equal one.

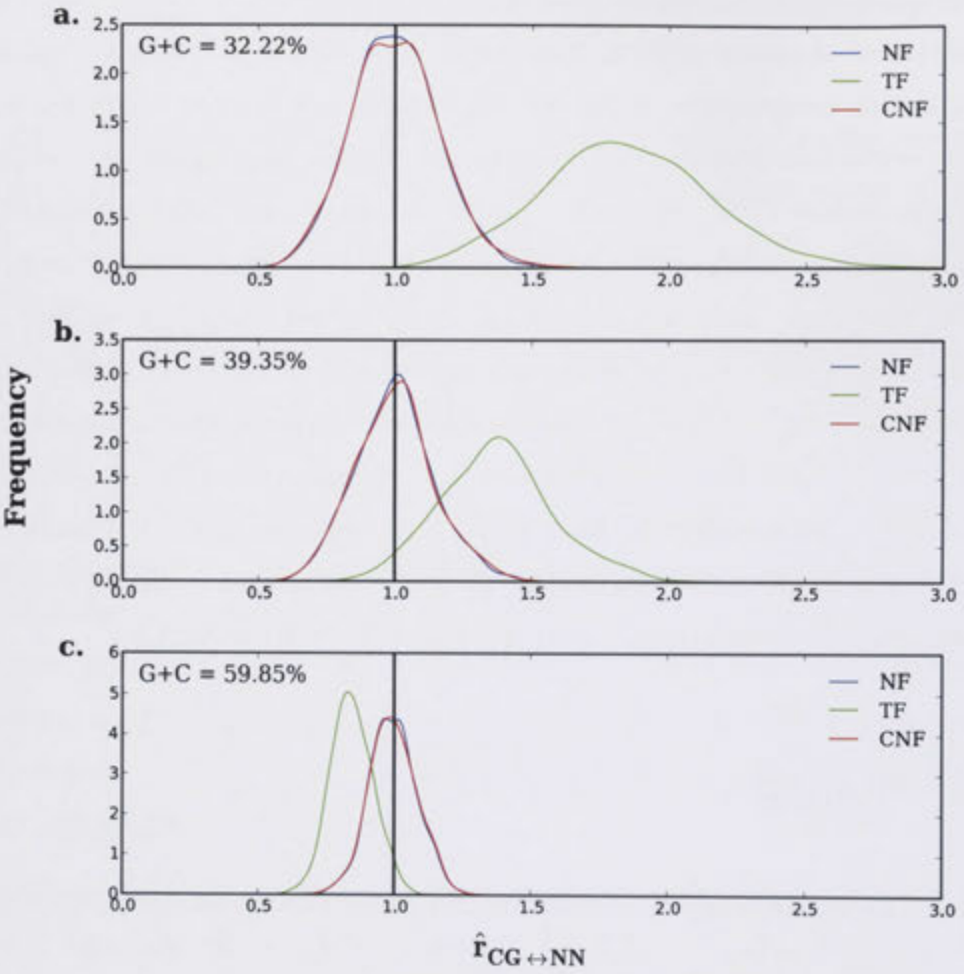

Figure 2.1: Comparison between model specifications of the estimated rate of CpG substitutions in context-free data. A GTR substitution model was fitted to three intronic alignments, with relatively **a.** low **b.** average and **c.** high G+C%. Using each set of fitted values, 1000 alignments were simulated, and $GTR + CG \leftrightarrow NN$ models fitted to the simulated alignments. Kernel density plots based on the distribution of $\hat{r}_{TG\leftrightarrow CG}$ estimates are shown. As the alignments were simulated according to an independent process, the expected value of $r_{TG\leftrightarrow CG}$ is 1.

The CNF and NF models did not detect any effect of the CpG context when applied to alignments simulated using a dinucleotide GTR substitution model with CNF motif frequencies. The TF model detected an effect of the CpG context that varied in strength and direction with the G+C% of the simulated alignments. Maximum likelihood estimates of $\hat{r}_{CG\leftrightarrow NN}$ for each model form are contrasted in Figure 2.1. As the alignments were simulated according to an independent nucleotide substitution process, the expected value of $r_{TG\leftrightarrow CG}$ is one. Estimates from the CNF and NF models were in close agreement and centered at approximately 1 for each set of alignments. Median estimates from the TF models varied with from > 1 for the lowest $G + C\%$ set, to < 1 for the highest $G + C\%$ set.

When applied to a real intronic alignment, the estimated effects of sequence context differed between model forms Figure 2.2). TF, CNF and NF variants of the 48 models including the GTR parameters and one additional dinucleotide parameter (the $GTR + dinuc$ models) were fitted to the alignment used for simulation in Figure 2.1 b. Unsurprisingly, given the rarity of CpG dinucleotides, estimated rates of substitutions involving CpG dinucleotides differed most between models. The TF model indicated the strongest effect of the CpG context on substitution rates, and the NF model indicated the weakest effect. Estimates from the NF model concurred more closely than the TF estimates to the CNF estimates when the CNF estimates indicated a weak context effect. The reverse was true when CNF estimates indicated a strong context effect. The CNF model found that CpG transitions were elevated by about 8-fold over transitions in general, and CpG transversions were elevated by about 4-fold over transversions in general (Figure 2.2 a. and b.). In contrast to the other model forms, the NF model did not detect a strong effect of context for CpG transversions.

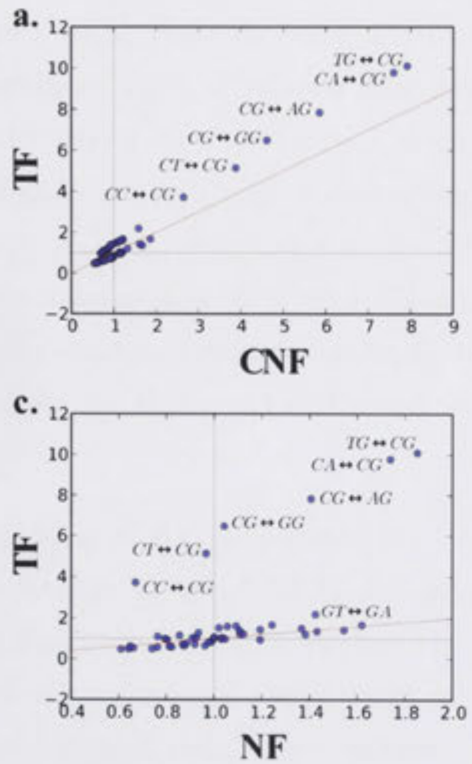

Figure 2.2: Comparison of dinucleotide parameter estimates between models. Dinucleotide parameters (*dinuc*) were estimated using a $GTR + dinuc$ substitution model for the conditional (CNF), dinucleotide (TF) and single nucleotide (NF) motif frequency specifications. Models were fitted an intronic alignment (ENSG00000115423) with a G+C% typical of the alignments in the intronic data set. Parameter estimates that differed by more than 0.75 between models are annotated with the substitution represented. The line where points on the X-axis and Y-axis are equal is shown in red, and black lines are drawn at the value 1 for each model. A parameter value of 1 is interpreted as the absence of a context effect.

Hereafter, all analyses use the CNF model form.

2.3.2 Influential dinucleotide contexts

The individual contributions of each dinucleotide parameter in improving model fit were assessed. The $GTR + dinuc$ models were fitted to the intronic data set. For each alignment in the data set, models were ranked according to the improvement in likelihood ratio over a GTR model. Top ranking dinucleotide parameters are shown in Table 2.2. Model rank was very heterogeneous between alignments, with all 48 models ranking within the top 10 for at least one alignment and 17 models ranking within the top 4 for at least one alignment. Unsurprisingly, the most influential parameters were transitions involving CpG dinucleotides. The next six most influential parameters were also transitions, involving dinucleotides composed of two A-T pairs.

For the dinucleotide parameters in Table 2.2, LR tests were performed to evaluate how much the parameter (denoted $dinuc$) improved the model fit over the GTR model; whether the $dinuc$ parameter was equally influential when the two most influential parameters ($r_{TG \leftrightarrow CG}{}^{*}$) were also included in the model; and whether $r_{TG \leftrightarrow CG}{}^{*}$ were equally influential when added to a $GTR + dinuc$ model. Pairs of strand complementary dinucleotide parameters are denoted with $*$, e.g. $r_{TG \leftrightarrow CG}{}^{*}$. Average likelihood ratio statistics for the intronic data set are presented in Table 2.3. Note that total likelihood improvement provided by a $GTR + TG \leftrightarrow CG^{*} + dinuc$ model over a GTR model does not depend on whether the $dinuc$ parameter or the $TG \leftrightarrow CG^{*}$ parameters are added to the model first, which can be seen by summing the relevant columns of Table 2.3.

Table 2.2: Most influential dinucleotide parameters for the intronic data set. Dinucleotide parameters were ranked by the extent to which likelihood improved when the dinucleotide parameter was added to a GTR model. Parameters with a top 6 rank for at least one alignment are ranked by the percentage of alignments in which they had a top 6 likelihood rank. The **Total** column shows the percentage of alignments for which a parameter ranked in the top 6.

$GTR+$	Rank (%)						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
$TG \leftrightarrow CG$	52.06	47.94	0.00	0.00	0.00	0.00	100.00
$CA \leftrightarrow CG$	47.94	52.06	0.00	0.00	0.00	0.00	100.00
$TT \leftrightarrow CT$	0.00	0.00	19.64	19.10	17.54	15.31	71.59
$AA \leftrightarrow AG$	0.00	0.00	17.84	20.28	18.44	14.52	71.08
$AT \leftrightarrow GT$	0.00	0.00	21.27	12.84	10.00	8.47	52.58
$AT \leftrightarrow AC$	0.00	0.00	21.09	12.05	10.00	8.56	51.70
$AA \leftrightarrow GA$	0.00	0.00	6.81	12.44	10.61	10.30	40.16
$TT \leftrightarrow TC$	0.00	0.00	7.02	10.64	10.85	10.52	39.02
$CT \leftrightarrow CG$	0.00	0.00	1.60	2.41	3.89	6.09	13.98
$CG \leftrightarrow AG$	0.00	0.00	1.42	2.02	3.83	5.27	12.53
$GA \leftrightarrow GG$	0.00	0.00	0.72	1.96	3.46	5.36	11.51
$TC \leftrightarrow CC$	0.00	0.00	0.99	1.93	3.77	4.61	11.30
$CG \leftrightarrow GG$	0.00	0.00	0.66	1.27	2.29	2.95	7.17
$CC \leftrightarrow CG$	0.00	0.00	0.36	1.05	1.87	3.28	6.57
$AG \leftrightarrow GG$	0.00	0.00	0.24	0.84	1.72	2.59	5.39
$CT \leftrightarrow CC$	0.00	0.00	0.33	1.14	1.63	1.99	5.09
$TA \leftrightarrow CA$	0.00	0.00	0.00	0.00	0.09	0.06	0.15
$TA \leftrightarrow TG$	0.00	0.00	0.00	0.03	0.03	0.06	0.12
$AT \leftrightarrow AG$	0.00	0.00	0.00	0.00	0.00	0.03	0.03
$TC \leftrightarrow TG$	0.00	0.00	0.00	0.00	0.00	0.03	0.03

The improvement in model fit obtained by including dinucleotide parameters in a model was dependent on the other parameters in the model. The influence of dinucleotide substitution parameters modeling $C \leftrightarrow T^*$ substitutions in non-CpG sequence contexts, for example $r_{TT \leftrightarrow CT^*}$, generally decreased when the $r_{TG \leftrightarrow CG}{}^*$ parameters were also included in the model, indicating that a portion of the influence of these parameters in improving model fit is due to the same cause. Surprisingly, the influence of the $r_{AT \leftrightarrow GT}{}^*$ parameters increased when the

Table 2.3: LR tests of dinucleotide importance for alignments in the intronic data set. The leftmost column '*dinuc*' indicates the dinucleotide parameter which was added either to the null and the alternate model (4th column), or just to the alternate model (2nd and 3rd columns). The remaining columns show the average improvement in model fit ($\bar{LR}$) for the alternate model over the null model, where column headings show the alternate model parameterisation over the null model parameterisation.

	$\bar{LR}\left(\frac{alternate}{null}\right)$		
$dinuc$	$\frac{GTR+dinuc}{GTR}$	$\frac{GTR+TG\leftrightarrow CG^*+dinuc}{GTR+TG\leftrightarrow CG^*}$	$\frac{GTR+dinuc+TG\leftrightarrow CG^*}{GTR+dinuc}$
$TG \leftrightarrow CG^*$	609.51	-	-
$TT \leftrightarrow CT$	39.83	19.51	589.19
$AA \leftrightarrow AG$	39.19	19.23	589.55
$AT \leftrightarrow GT$	34.72	59.24	634.03
$AT \leftrightarrow AC$	33.80	58.05	633.76
$AA \leftrightarrow GA$	27.96	12.66	594.21
$TT \leftrightarrow TC$	27.35	12.18	594.33
$CT \leftrightarrow CG$	18.79	21.81	612.53
$CG \leftrightarrow AG$	18.19	21.13	612.45
$TC \leftrightarrow CC$	17.37	7.20	599.33
$GA \leftrightarrow GG$	17.10	7.04	599.45
$CG \leftrightarrow GG$	14.65	16.51	611.37
$CC \leftrightarrow CG$	14.37	16.25	611.39
$AG \leftrightarrow GG$	12.17	3.75	601.09
$CT \leftrightarrow CC$	11.54	3.61	601.57
$TA \leftrightarrow TG$	5.19	3.22	607.54
$TA \leftrightarrow CA$	5.04	3.09	607.55

$r_{TG\leftrightarrow CG^*}$ parameters were also included in the model, suggesting a synergistic effect. Whether this has a biological cause, or is a feature of the model formulation is unclear.

2.4 Discussion

In reversible, context-dependent substitution models, entries of the instantaneous rate matrix Q are typically weighted by the frequency of the motif resulting from the substitution. Here it was shown that the definition of these motifs affects estimates of context-dependent substitution rates. One of the two most commonly used motif frequency definitions (TF) was shown here to infer that substitution is context-dependent in simulated data that evolves with no context effects but non-multiplicative dinucleotide frequencies. The equivalent models for a codon alphabet both incorrectly infer context-effects when none exist, with the extent to which estimates deviate from expectations dependent on the motif frequency composition of the sequence being analysed (Yap et al., 2010). As in Lindsay et al. (2008), the NF model did not detect an influence of the CpG context in simulated context-free data in this study.

The ability to accurately model the biased composition of real DNA sequences is a crucial feature of a context-dependent substitution model. This analysis has shown that, as in Yap et al. (2010), the NF model performs poorly when detecting context effects in real data. The CNF and TF models both detected a strong influence of context for CpG transversions. These estimates are in agreement with analyses of human, disease-causing mutations, which also indicate that CpG transversions occur at an elevated rate (e.g. Krawczak et al., 1998). When compared with the TF and NF models, the CNF model is the only model to perform equally well when detecting the presence and the absence of context effects in non-coding data. The equivalent result has been previously shown for coding data (Yap et al., 2010).

Motif frequency specification had little impact on the dinucleotide parameters identified as the most important for improving model fit. The dinucleotide contexts estimated to be the most consistently influential using the CNF model were in general agreement with the influential contexts identified using the TF model in Lindsay et al. (2008). The NF model did not detect this effect for the intronic alignments considered. These analyses suggest that the TF model is prone to type I error, whilst the NF model is prone to type II error.

Chapter 3

Regional substitution process variation

Previous studies have assumed that male-bias is consistent across the genome and that regional variation in the substitution process arises from non-sex-biased processes. If this assumption is true, failure to consider regional processes when estimating male-bias will confound results. In this chapter, I examine regional substitution rate variation in relation to male-bias and consider how best to disentangle the influences of sex-biased from non-sex-biased processes on the substitution rate. I demonstrate that that a simple strategy of matching alignments according to their compositional properties or propensity to undergo recombination does not result in meaningful estimates of male-bias (α) for the data set under consideration in this study.

3.1 Introduction

Substitution rate heterogeneity poses a fundamental challenge for the estimation of male-bias, as differences in substitution process arising from regional factors must be distinguished from differences arising from sex. Nucleotide substitution rates vary considerably within chromosomes, between autosomes (Lercher et al. (2001)), and between autosomes and sex chromosomes. Estimates of male bias are sensitive to genomic location of the sequences compared (Berlin et al., 2006, Smith and Hurst, 1999). A further complication is that Miyata et al.'s (1987) α statistic assumes that the relative time each chromosome spends in the male and female germ cell environments determines the extent to which it experiences male-biased mutation. As such, Miyata et al.'s (1987) α is unsuited for measuring sex-biased processes that are not time-dependent.

Several previous analyses have attempted to eliminate the confounding effects of regional substitution rate variation by estimating sex-bias for duplicated sequences (e.g. Goetting-Minesky and Makova, 2006, Bohossian et al., 2000, Chang and Li, 1995). As duplicated sequences share an ancestral sequence, members of a family of duplicated sequences should, at least initially, be affected by the same mutation biases caused by local sequence context. However, Berlin et al. (2006) found no correlation between the substitution rates of homologous, non-recombining introns shared by the sex chromosomes in birds, indicating that duplicated sequences are not immune from regional substitution process biases (Berlin et al., 2006).

An alternative approach to dealing with regional substitution rate variation when estimating male-bias has been to "correct" substitution rate estimates for effects of sequence composition and recombination rate using regression techniques (Taylor et al., 2006). However, this study did not examine whether the assumptions underlying the α statistic for estimating male-bias were better satisfied once the correction was applied.

The availability of whole genome sequences, including the recently published macaque genome (Gibbs et al., 2007), enables the relationship between regional substitution rate variation and male-biased mutation to be examined on a large scale. Of particular interest is whether the assumptions of Miyata et al.'s (1987) can be met with appropriate matching of data for regional influences, and whether there is evidence for a single, genome-wide male-bias. Amongst the most favoured candidate causes of regional substitution rate variation are nucleotide composition, recombination, and telomere-specific effects. The proposed mechanisms by which each of these factors are thought to influence nucleotide substitution rates are introduced below.

Substitution rate heterogeneity coincides with variation in nucleotide composition (e.g. Bielawski et al., 2000, Webster et al., 2003). The genomes of warm-blooded vertebrates are comprised of large blocks of relatively homogeneous nucleotide composition, called isochores, typically each at least 300 kb long (Bernardi, 1993). The nucleotide composition of a sequence is typically measured as the frequency of guanine and cytosine nucleotides (G+C%).

Sequence composition itself has been proposed to cause of substitution rate heterogeneity. In comparison with A:T nucleotide pairs, G:C pairs have greater thermodynamic stability. Certain potentially mutagenic nucleotide alterations, such as the deamination of cytosine to uracil or mC to thymine, occur much more readily in single-stranded than in double-stranded DNA (Lindahl and Nyberg, 1974). As more energy input is required to separate a G:C pair then an A:T pair, regions with an elevated G+C% may undergo less transient, localised DNA strand separation (DNA breathing) than regions with a high A+T%, and consequently have a slower rate of mC mutation. Another compositional feature that may have a large impact on substitution rate and process variation is the frequency of CpG dinucleotides. As mutations at CpG dinucleotides occur at a highly elevated rate (Cooper and Youssoufian, 1988), differences in the CpG dinucleotide frequency are expected to affect the substitution rate.

The most popular hypothesis to account for isochore structure, known as the BGC hypothesis (see page 25), is that nucleotide mismatches generated by recombination are subject to biased repair favouring G and C nucleotides (Galtier et al., 2001). Consequently, regions with an elevated recombination rate are expected to develop an elevated G+C% content. The BGC hypothesis is motivated by the observation that mismatches of a G or C nucleotide with an A or T nucleotide are corrected with a bias favouring the G or C in monkey kidney cells (Brown and Jiricny, 1988). As estimates of the BGC rate are not easily obtained, in previous studies the recombination rate has been used to approximate the rate of BGC (e.g. Duret and Arndt, 2008). The location of a sequence with respect to the nearest telomere can also be used as a crude proxy measure of the male-specific rate of BGC, as the male recombination rate is greatest near telomeres (Kong et al., 2002). Telomere-specific changes in the substitution process consistent with BGC have been previously identified (Arndt et al., 2005), including changes indicative of sex-differences in the effect of recombination on the mutation process Dreszer et al. (2007).

Mutation bias hypotheses for isochore structure attribute nucleotide compositional heterogeneity to the regionally biased occurrence of DNA damage and repair. In eukaryotic cells, DNA sequences are extensively coiled around proteins to form chromatin. The extent of coiling and therefore the accessibility of sequences to mutagens and repair systems varies with nucleotide composition

(Gong et al., 2005, Filipski, 1987, Ehrenhofer-Murray, 2004). Chromatin compactness is inversely correlated with substitution rate (Prendergast et al., 2007), and periodic variations in the human substitution process in accordance with known histone locations have been observed (Ying et al., 2010). Another mutation-bias explanation for substitution rate heterogeneity, which is more compatible with the replication origin hypothesis for male bias than other explanations, is that the spectrum of replication errors changes throughout the replication cycle as a result of changes in the pool of free nucleotides (Wolfe et al., 1989). Support for this hypothesis comes from observations that the spectrum of errors made by DNA polymerase is affected by the concentrations and relative proportions of the nucleotide percursors, which change throughout the replication cycle (Martomo and Mathews, 2002).

In this chapter, the uniqueness of the substitution process of X-linked loci is considered in the context of genome-wide substitution process heterogeneity, and the extent to which substitution rate differences between X-linked and autosomal alignments truly reflect sex- rather than regional effects evaluated. The relationships between nucleotide substitution rate and two direct measures of sequence composition (G+C% and CpG frequency) and several measures that may be correlated with the BGC rate (male and female-recombination rate estimates and telomere proximity) are assessed for X-linked and autosomal alignments. The influence of these factors on the single nucleotide substitution process was also assessed. Analyses were performed using the intergenic data set and, as the recombination rate estimates used were human-specific, analyses of substitution rates were restricted to the human rate estimates.

3.2 Methods

3.2.1 Estimation of recombination rates

Human sex-specific recombination rates were obtained from the deCODE database (Kong et al., 2002). For the majority of alignments in the untranscribed data set, no recombination rate estimate was available in the region spanned by the alignment. Splines were fitted to graphs of recombination rate estimates versus chromosomal location using the pspline package in R, with settings adjusted so that the fitted spline passed through all estimates. Goodness-of-fit was verified by visual inspection. Sex-specific recombination rates were extrapolated from the splines for the start coordinate of each alignment. (Due to the removal of gaps and low complexity sequences, the genomic coordinates of the alignments were not reliable indicators of the central position.) Recombination rates were not extrapolated for alignments located more distally than the most distal recombination rate estimate, or for locations with at least 5Mb between consecutive recombination rate estimates. The latter criterion was chosen specifically to remove centromeric alignments, for which recombination rate estimates were unavailable.

3.2.2 Estimation of distance to telomere

The distance of each alignment from the closest telomere was calculated, using the estimates of total chromosome length published by Lander (2001). As for the estimation of recombination rates, distances were estimated with respect to the start of the alignment.

Estimation of confidence intervals for α

Confidence intervals for α estimates were approximated by resampling alignments with replacement from a data set and calculating α using the ratio of the sample mean substitution rate estimates for the X-linked and autosomal data sets. 1000 samples of X-linked and autosomal alignments were taken, each sample consisting of the same number of estimates as the corresponding observed data set. 95% confidence intervals were inferred from the distribution of sample α statistics.

3.2.3 Sliding window comparison of α confidence intervals

Confidence intervals were estimated for the relationships between the substitution rate predictors and α estimates by partitioning substitution rate estimates according to the value of the predictor, and constructing confidence intervals for partitions. Confidence intervals were constructed as described above with the exception that resampling was performed only 500 times for each partition. For each predictor, partition size was determined by dividing the range of the predictor variable by five, where range is defined as the union of the ranges for the X-linked and autosomal alignments. Starting from a window centered on the lower endpoint of the range, confidence intervals were calculated for overlapping partitions. Successive partitions started at the central point of the previous partition. For example, if the ranges of the predictor variable for the X-linked and autosomal alignments were (0, 20) and (5, 25) respectively, the range of the predictor would be (5, 20), the partition size would be 3, and confidence intervals would be calculated for the ranges $(4, 6), (5, 7),\ldots(19, 21)$.

3.3 Results

3.3.1 Dinucleotide composition of X-linked and autosomal alignments

Compositional differences between X-linked and autosomal alignments may contribute to their branch length differences. As different combinations of dinucleotide frequencies can result in the same G+C%, the suitability of G+C% as an indicator of dinucleotide motif composition was examined. Wilcoxon signed-rank tests were performed to evaluate whether X-linked and autosomal sequences had the same dinucleotide frequency composition. A p-value of 0.05 was chosen to nominally indicate significance. Figure 3.1 indicates the comparisons that were significant after a Bonferroni correction was applied to adjust the p-value cutoff to account for multiple testing. The corresponding G+C% distribution for each chromosome is also shown. Intergenic, X-linked alignments tended to have a lower frequency of CpG dinucleotides than autosomal alignments of similar G+C%, but the frequencies of all other dinucleotides generally did not differ.

3.3.2 Predictors of variation in human substitution rate variation

The effect of matching X-linked and autosomal alignments according to their composition, telomeric location or recombination rate, and estimating α for matched alignments was evaluated. Substitution rate estimates were first plotted, and trend functions calculated (Figure 3.2). The trend functions were used to derive α estimates as a function of each predictor variable. Bootstrap 95%

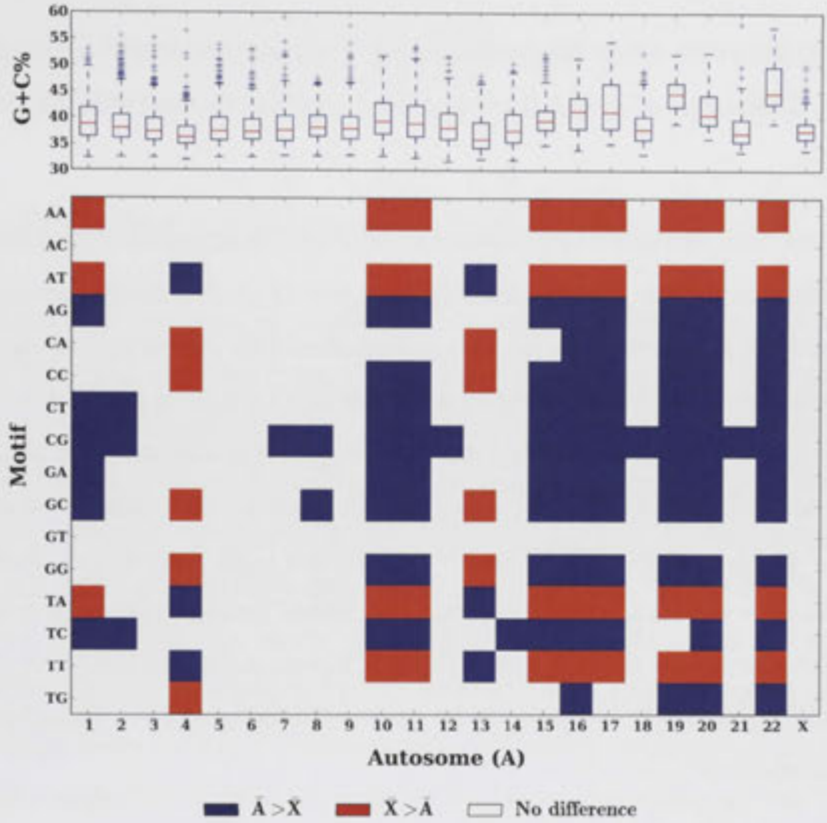

Figure 3.1: Comparison of G+C% and dinucleotide motif probabilities between intergenic X-linked and autosomal data. The top panel shows the G+C% distribution of intergenic alignments, sorted by the chromosomal location of the human sequences. Whiskers show the G+C% range for alignments with G+C% within (1.5 times the interquartile range) of the upper and lower quartiles. Alignments with G+C% outside this range are indicated with + signs. The lower panel summarises the results of Wilcoxon signed-rank tests of whether the distribution of each dinucleotide motif in X-linked alignments differed from the distribution of the same motif in each of the autosomes. A p-value cutoff of 0.05 was chosen to indicate significant difference. Tests that were significant after a Bonferroni correction for multiple testing was applied are indicated in colour. Blue squares indicate tests with a significant p-value where the mean dinucleotide frequency on the autosome was greater than that on the X chromosome ($\bar{A} > \bar{X}$), and red squares indicate significant tests where the mean dinucleotide frequency was greater for the X-linked alignments ($\bar{X} > \bar{A}$).

confidence intervals were estimated using a sliding window technique (see Methods). Estimates of α and the associated confidence intervals are shown in Figure 3.3.

Substitution rate differences between X-linked and autosomal alignments were not consistent across the genome for alignments with similar nucleotide composition (Figure 3.2, a. and b.). The human substitution rate of autosomal alignments tended to increase with increasing G+C% and CpG frequency, but the opposite was true of X-linked alignments. Unsurprisingly, estimates of α for alignments matched according to their G+C% or CpG frequency were highly variable. In Figure 3.3 a. (α versus G+C%) and b. (α versus CpG frequency), there is an asymptote as α estimates tend towards ∞. These invalid values of α indicate that the remaining substitution rate difference between the matched X-linked and autosomal alignments is too great to be accounted for only by a time-dependent male mutation bias.

Matching alignments according to their female recombination rate resulted in a more consistent relationship between the substitution rates of X-linked and autosomal alignments (Figure 3.2 c.). Estimates of α and their confidence intervals for alignments matched according to their female recombination rate remained too variable to be meaningful Figure 3.3 c.. The positive relationship between the male recombination rate and the autosomal substitution rate (Figure 3.2 d.), and the high substitution rate of some autosomal, telomeric alignments (Figure 3.2 e.) both suggest that recombination in males contributes to the substitution rate. Simultaneously correcting for effects of male and female recombination was not attempted, as it is not straightforward to determine how the male and female rates should be weighted, and such a correction was not required for subsequent analyses.

3.3.3 Predictors of substitution process variation

The influence of the factors considered above in relation to the human substitution rate was evaluated for the rescaled GTR parameter estimates using Spearman's rank correlation test. Spearman's ρ statistics for the 60 separate tests are shown in Table 3.1. The G+C% was the strongest predictor of the substitution process of both autosomal and X-linked data. Correlations between G+C% and GTR parameter estimates were weaker for X-linked data than for autosomal data. CpG frequency was a significant predictor of the autosomal substitution process, but less so for the X-linked data. Although other significant relationships between GTR estimates and alignment properties were detected, the small ρ statistics in Table 3.1 indicate that these associations were generally weak. The one exception was X-linked $\hat{r}_{T \leftrightarrow A}$ estimates, which varied significantly with female recombination rate and telomeric location, but not with G+C%.

GTR parameters estimates varied linearly with G+C% for autosomal sequences (Figure 3.4). The relative proportion of transitions increased with G+C%. Of the transversion terms, the proportion of $\hat{r}_{C \leftrightarrow G}$ substitutions decreases most rapidly with G+C%. As indicated by the Spearman tests, there was little relationship between the $\hat{r}_{T \leftrightarrow A}$ estimates and G+C%. Transition parameters estimates were highly variable even amongst alignments with the same G+C%.

Table 3.1: Association between GTR parameter estimates and sequence features. Spearman's rank correlation statistics were calculated to test for associations. Spearman's ρ statistic is shown for each comparison to 2 decimal places, and associations that were significant at the 0.001 (**) or 0.05 (*) level, after the p-value cutoff was adjusted to correct for the multiple comparisons, are indicated. GTR parameter estimates were rescaled to sum to one, and were estimated for intergenic X-linked (X) and autosomal (A) data using the CpG baseline model. Abbreviations: ♀ - Female ♂ - Male **rr** - recombination rate.

	Data	$\hat{\mathbf{r}}_{\mathbf{A}\leftrightarrow\mathbf{G}}$	$\hat{\mathbf{r}}_{\mathbf{T}\leftrightarrow\mathbf{C}}$	$\hat{\mathbf{r}}_{\mathbf{C}\leftrightarrow\mathbf{G}}$	$\hat{\mathbf{r}}_{\mathbf{T}\leftrightarrow\mathbf{G}}$	$\hat{\mathbf{r}}_{\mathbf{C}\leftrightarrow\mathbf{A}}$	$\hat{\mathbf{r}}_{\mathbf{T}\leftrightarrow\mathbf{A}}$
G+C%	A	0.41 **	0.36 **	−0.72 **	−0.45 **	−0.45 **	0.25 **
	X	0.35 **	0.18 *	−0.58 **	−0.44 **	−0.38 **	0.05
CpG frequency	A	0.35 **	0.24 **	−0.54 **	−0.34 **	−0.34 **	0.15 **
	X	0.24 *	0.08	−0.19 *	−0.27 **	−0.25 *	−0.13
Distance to telomere	A	−0.10 **	−0.06 **	0.16 **	0.07 **	0.07 **	−0.03
	X	−0.08	−0.13	−0.04	0.20	0.15	0.40 **
♀ rr	A	0.04	0.02	−0.07 **	−0.01	−0.02	−0.00
	X	−0.02	0.12	0.11	−0.11	−0.00	−0.31 **
♂ rr	A	0.05 **	0.01	−0.06 **	−0.02	−0.03	−0.03

3.4 Discussion

Matching X-linked and autosomal alignments according to their nucleotide or CpG composition was not sufficient to enable meaningful α estimates to be calculated. Although the substitution rate difference between X-linked and autosomal alignments was relatively constant for alignments matched according to their recombination rate in the female germline, α estimates remained very sensitive to fluctuations in substitution rate estimates for X-linked data. Matching alignments for several features might generally decrease the variance of α estimates, but reducing the volume of data considered could have the opposite effect. Considering the confidence intervals for α estimated in here, it is no surprise that previous studies have reported a wide range of α statistics even when attempts have been made to eliminate the confounding regional effects on nucleotide substitution rates.

The replication origin hypothesis assumes that the majority of mutations are caused by replication errors, and consequently implicitly assumes that regional substitution rate heterogeneity is also caused by replication errors. Although the spectrum of replication errors is believed to change throughout the replication cycle as a result of changes in the composition of the free nucleotide pool (Wolfe et al., 1989, Martomo and Mathews, 2002), the male-to-female ratio of replication cycles remains constant. Therefore, the replication origin hypothesis predicts that male bias should be constant amongst sequences that replicate at the same time. At least in somatic cells, replication timing is strongly correlated with sequence composition (Schmegner et al., 2007, Woodfine et al., 2004). However, in this study α estimates increased with G+C% to invalid values, indicating that regional substitution rate heterogeneity is not fully explained by changes in the frequency of replication errors over the course of a replication cycle. A recent analysis of the relationship between rodent substitution rates and replication timing in embryonic cells reported a similar finding (Pink and Hurst, 2009). Changes in replication timing associated with X-inactivation are not likely to have contributed extensively to this result, as random X-inactivation is reversed early during the germ cell cycle (Morgan et al., 2005).

An influence of recombination on the human substitution rate may explain why the relationships between sequence composition and the rate estimates for X-linked and autosomal alignments differed. If the female recombination rate affects the nucleotide substitution rate, the male recombination rate is also likely to be influential. Indeed, several studies indicate that recombination has more effect on the substitution rate and process when it occurs in the male germline than the female germline (Webster et al., 2005, Dreszer et al., 2007, Duret and Arndt, 2008), and in this study the gradient of the relationship between the male recombination rate and human substitution rate of autosomal alignments was greater than that for the female recombination rate. The male recombination rate is greatest, and also greatest with respect to the female recombination rate, for sequences located close to a telomere (Broman et al., 1998, Kong et al., 2002). Such sequences also tend to be G+C-rich. As X-linked alignments do not recombine in males, except at the pseudo-autosomal region, differences between X-linked and autosomal alignments located close to a telomere may be indicative of the influence of the male recombination rate on the substitution rate. Recombination could therefore explain all of the relationships between substitution rates and sequence features observed in this study: if male recombination increases the substitution rate of G+C-rich autosomal but not X-linked sequences, no telomere effect is expected for X-linked data and a stronger relationship between the female recombination rate and the nucleotide substitution rate of X-linked than autosomal data is predicted.

Although the substitution rate was correlated with the distance of alignments from a telomere and their male and female recombination rates, the GTR parameters were not similarly correlated except in a small number of cases. The reversible models used in this analysis are not suited for detecting the directional effects of BGC, i.e. an increase in the rate of substitutions that create a G or C compared with the rate of the reverse substitutions. The GTR parameter estimates were, however, strongly correlated with the G+C% and CpG frequency. The elevated CpG transversion and transversion rates may contribute to these observations. The only type of single nucleotide substitution that cannot either occur at or create a CpG dinucleotide is $T \leftrightarrow A$, and $\hat{r}_{T\leftrightarrow A}$ estimates were the only instantaneous substitution rate estimates to vary significantly with the female recombination rate.

The results presented here suggest that the factors that cause regional substitution rate variation also contribute to sex-bias. If the converse was true, estimates of male bias should be consistent for alignments matched according to regional influences. Matching X-linked and autosomal alignments according to their female recombination rate resulted in the most consistent estimates of male-bias. However, if female recombination contributes to substitution rate heterogeneity, it is reasonable to assume that male recombination rate does also. Any influence of male recombination rate on the substitution rate is expected to affect estimates of sex-bias, for the reasons discussed above. Another consideration is that G+C% was the best predictor of the single nucleotide substitution rate spectrum of both X-linked and autosomal alignments, whereas the female recombination rate varied with substitution rate but not process. Although G+C% therefore the best predictor of regional variation in the substitution process, estimates of male bias increased with increasing G+C%. Consequently the results of this chapter do not support the existence of a single, genome-wide male-bias, although individual contributing factors may have a consistent effect across the genome. Motivated by the results presented in this chapter, the approach taken for estimating the contribution of methylation to male-bias in the following chapter is to compute genome-wide mean estimates of male-bias, and also consider how estimates vary with nucleotide composition.

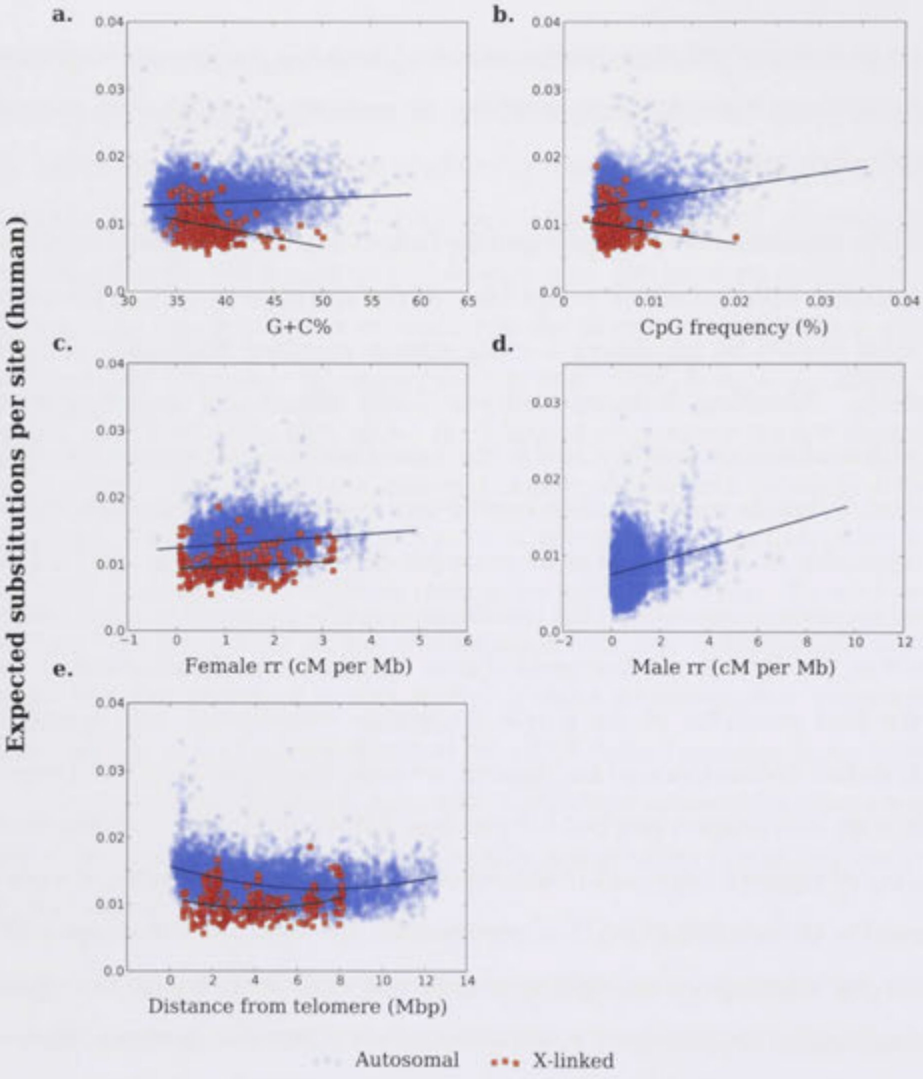

Figure 3.2: Human substitution rate estimates versus rate predictors. **a.** G+C% **b.** CpG frequency **c.** Female-specific recombination rate (rr) **d.** Male-specific recombination rate **e.** Distance from the human telomere. G+C% and CpG frequency are the average values across an entire alignment. Polynomials of degree 2 were fitted to each data set for **e.**, and degree 1 polynomials fitted to the other figures. Fitted values are plotted. Each data point represents the maximum likelihood estimate of the human substitution rate for one of the alignments in the intergenic data set. Abbreviations: **rr** recombination rate **cM** centiMorgan **Mb** Megabase **bp** base pairs

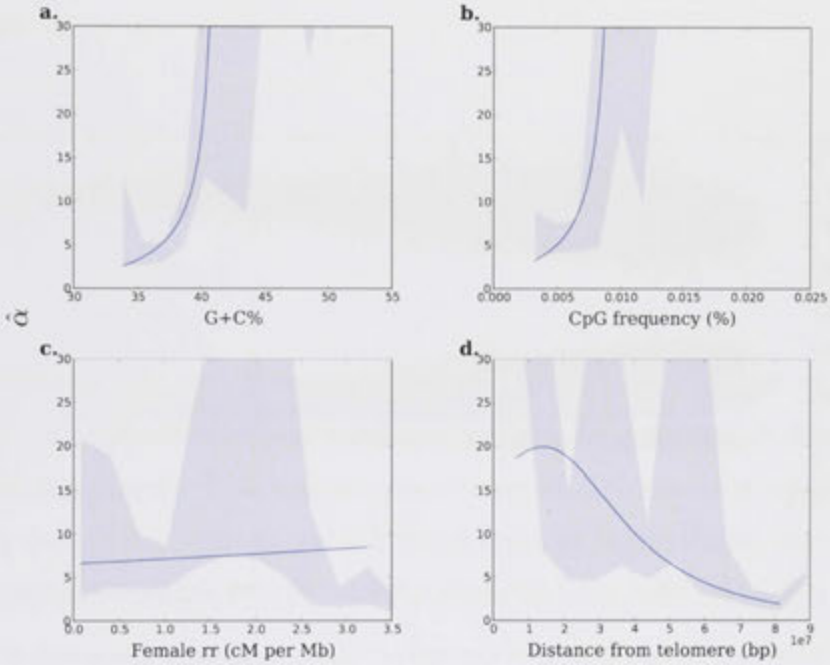

Figure 3.3: Human-specific $\hat{\alpha}$ versus alignment features. **a.** G+C% **b.** CpG frequency **c.** Female-specific recombination rate (rr) **d.** Distance from the human telomere. $\hat{\alpha}$ was estimated using the trend functions in Figure 3.2. Approximate 95% bootstrap confidence regions are shaded. Abbreviations are defined as for Figure 3.2.

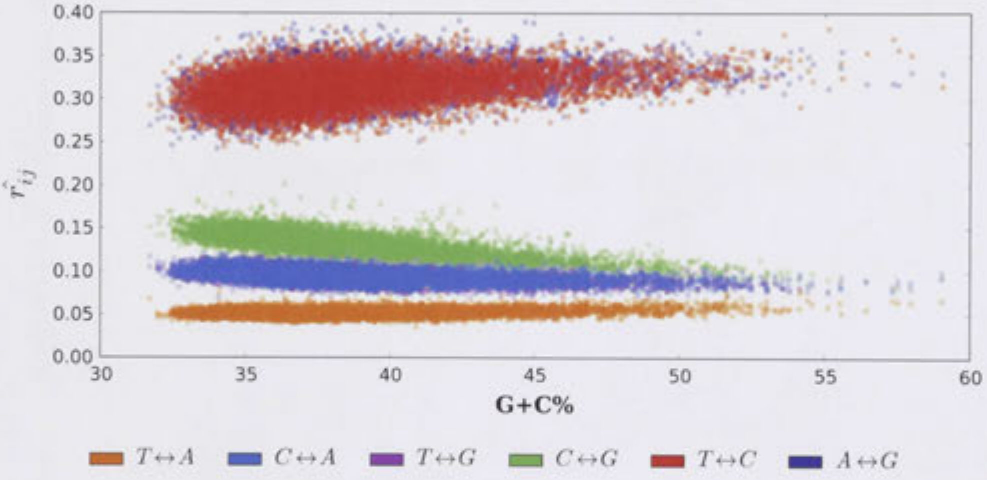

Figure 3.4: Variation in the substitution rate spectrum of intergenic, autosomal data with G+C%. Relative rate estimates were estimated with a GTR substitution model and are scaled to sum to 1. The x-axis shows the G+C% of the alignment.

Chapter 4

Methylation

The hypothesis that sex differences in the frequency of mutation at mC nucleotides are a major cause of male-biased evolution is evaluated in this chapter. I consider whether the CpG transition and transversion rates are equally male-biased in comparison with the rates of other types of substitutions, and estimate the proportional contribution of methylation to male biased-evolution.

4.1 Introduction

DNA methylation, which is the covalent addition of a methyl group to the 5′ position of a cytosine nucleotide, is one of the most important epigenetic modifications in mammals. It is involved in biological processes associated with chromatin condensation, including silencing retroelements, transcriptional regulation, and setting up tissue-specific and developmental stage-specific gene expression patterns. In mammals, most methylation occurs at cytosine nucleotides that have guanine as their 3′ neighbour, i.e. CpG dinucleotides.

Methylation occurs at 60-80% of the CpG dinucleotides in mature human germ cells (Trasler, 2006), the majority of which are located in repetitive sequences (Trasler, 2009). Many mammalian gene promoters include CpG-rich regions, known as CpG islands, that generally remain unmethylated. Outside CpG islands, the CpG dinucleotide is considerably underrepresented, presumably because of the high propensity of mC to undergo mutation (Coulondre et al., 1978).

Mutations of CpG dinucleotides are a major contributor to human genetic disease and evolution. Approximately one third of all human point mutations are transitions at CpG dinucleotides (El-Maarri et al., 1998), and 37% of human disease-causing point mutations occur at CpG dinucleotides (Cooper and Youssoufian, 1988). Transitions at CpG dinucleotides are estimated to occur at least an order of magnitude faster than that of other types of substitutions (Nachman and Crowell, 2000, Cooper and Youssoufian, 1988). Because of the highly elevated mutation rate of CpG dinucleotides, sex differences in germline biology that affect the mutation rate of CpG dinucleotides potentially have large effects on the male-bias in the mutation rate.

Mutations at mC nucleotides are not generally thought to be a consequence of replication errors (e.g. Taylor et al., 2006), although replication does affect the mC mutation rate (Lieb and Rehmat, 1997, see below). A key argument in favour of this hypothesis is that, unlike other types of substitutions, transitions at CpG dinucleotides accumulate at a constant ("clock-like") rate over time (Hwang and Green, 2004, Kim et al., 2006). If the rates of CpG and other types of substitutions are affected by different processes, they may also differ in male-bias. In the following sections, the processes that influence the rate of CpG substitution, and their likely activity in the germline are considered.

4.1.1 Determinants of the substitution rate of CpG dinucleotides

Whilst the factors that modulate the CpG substitution rate are not completely understood (Pfeifer, 2000, El-Maarri et al., 1998, Radford and Lobachevsky, 2008), the elevated mutation rate of methylated compared to unmethylated cytosine has been clearly demonstrated. Cytosine residues that are mutation hotspots when methylated in transgenic *Escherichia coli* are not hotspots in strains lacking DNA methylation (Coulondre et al., 1978). Mutations at mC nucleotides arise as a consquence of spontaneous hydrolytic deamination of mC, which produces the naturally occurring base thymine. The resulting G:T lesions are thought to cause difficulties for repair systems, so that transition mutations arise either at replication or following inaccurate repair. Deamination is widely accepted as the primary mechanism by which CpG mutations arise, although other processes can also cause lesions specifically at mC nucleotides (e.g. Pfeifer, 2000), and the deamination explanation does not account for the elevated rate of CpG transversions (Huttley, 2004, Siepel and Haussler, 2004, Blake et al., 1992).

The mutation rate of a CpG dinucleotide may be affected by its methylation status, the rate of mC deamination, the frequency and accuracy with which T:G lesions are repaired, and the chance of lesions being fixed as a mutations at replication before repair can take place. Mutations that occur for reasons other than deamination of mC will also contribute to the total mutation rate. Of these processes, deamination is often assumed to be the stage that most influences the mutation rate (e.g. Shen et al., 1994, Taylor et al., 2006, Fryxell and Moon, 2005).

The frequency of CpG transitions in humans is much less than the theoretical maximum that can be explained by deamination events if the mC deamination rate in humans is similar to the rate in *E. coli* (Shen et al., 1994). However, the difference between the rate of cytosine transitions in CpG and non-CpG contexts is much greater than the difference in the deamination rates of C and mC (Lindahl and Nyberg, 1974, Pfeifer, 2000), indicating that other factors influence the CpG transition rate.

The balance between the rates of DNA replication and repair also potentially contributes to the mutation rate difference between transitions within and outside the CpG dinucleotide context. Mismatches that arise from replication and from deamination of mC must be dealt with differently. For a replication error, the nascent strand must be corrected; whereas for a G:T mismatch arising from deamination of mC the T should always be reverted to a C regardless of strand. At least two different repair systems can correct G:T mismatches in eukaryotic cells. In addition to the MMR system which corrects DNA replication errors, eukaryotes have a dedicated repair enzyme, thymine glycosylase, that acts as part of the BER system to specifically excise the T from a G:T mismatch in the context of a CpG dinucleotide. The repair of G:T lesions by BER is biased in favour of retaining the G (Brown and Jiricny, 1988). In hamster ovary cells, MMR and BER compete to repair G:T lesions (Bill et al., 1998). Repair mediated by thymine glycosylase is much less efficient. As MMR does not share the correction bias of thymine glycosylase, a mismatch resulting from deamination of mC can be fixed as a mutation if it is repaired by the incorrect system. Intuitively, this scenario is most likely to occur when MMR is active, i.e. in cells that frequently replicate. Indeed, in *E. coli*, an mC site was found to only be a mutation hotspot in dividing cells (Lieb and Rehmat, 1997). Repair of G:T lesions may be targeted to actively transcribed genomic regions, as thymine glycosylase interacts with a variety of transcription factors (reviewed in Cortazar et al., 2007).

4.1.2 Methylation pattern changes during gametogenesis

Genome-wide changes in methylation patterns are crucial for the mammalian embryo to regain totipotency (Trasler, 2006, Morgan et al., 2005). The majority of DNA methylation is erased during early embryogenesis and reacquired during gametogenesis at a gender-specific time. Demethylation of the paternal and maternal genomes occurs via different mechanisms. The paternal genome undergoes active, i.e. enzyme-mediated, demethylation as the protamines are replaced by histones prior to the first DNA replication (Morgan et al., 2005, Trasler, 2006). As the maternal genome undergoes a more gradual demethylation process during the blastocyst stage, the paternal genome is hypomethylated in comparison with the maternal genome during the very early stages of embryogenesis.

Reacquisition of methylation patterns during gametogenesis is also gender-specific both in terms of timing and extent. Methylation of male germ cells occurs prenatally for most sites (Trasler, 2006), whilst methylation of oocytes occurs progressively during the oocyte growth phase around puberty and is completed by metaphase II of meiosis (Lucifero et al., 2002). The mature mammalian spermatozoa and oocyte have considerably different methylation patterns (Allegrucci et al., 2005), which are not limited to imprinted loci (Swales and Spears, 2005). Mature oocytes are globally hypomethylated compared to sperm, though both types of germ cells are less methylated than somatic tissues (Monk et al., 1987, Allegrucci et al., 2005, Driscoll and Migeon, 1990). In primates, the duration of time for which male germ cells are more methylated than female germ cells is much longer than the early embryonic period when the paternal genome is comparatively hypomethylated.

4.1.3 Potential consequences for male-biased evolution

The possibility that most CpG mutations are of paternal origin was suggested by Driscoll and Migeon (1990) after they observed that human oocytes are unmethylated compared to spermatocytes when they begin meiosis. Lower methylation levels, later acquisition of methylation and less frequent replication in oocytes suggest that CpG transitions should be male-biased. The CpG transition rate may also be affected by sex differences in germline transcription activity (discussed in Chapter 5) and differences in deamination rate caused by regional nucleotide composition.

As is typical of α estimates in general, estimates of the male-bias in the CpG mutation and substitution rates are highly variable (e.g. Ketterling et al., 1993, Becker et al., 1996, Taylor et al., 2006). McVean (2000) found that the very low human α estimate published by Bohossian et al. (2000) can be attributed to the low CpG frequency of the data analysed, assuming that male-bias at other sites derives from a strong bias at CpG sites. Ketterling et al. (1993) found the rate of hemophilia-causing CpG transitions was greater than the rate of non-CpG transitions, as was the rate of CpG compared to non-CpG transversions. Dinucleotide frequency differences between chromosomes also support a contribution of methylation to male-biased mutation (Huttley et al., 2000). However, estimates of the male-bias derived from substitution rate estimates do not consistently support a strong male-bias for CpG substitutions relative to other types of substitutions. The contribution of CpG substitutions to the male-biased substitution rate is typically estimated by dividing the sites of an alignment into "CpG" and "non-CpG" classes, and estimating the substitution rate separately for each class. Using this method, Smith and Hurst (1999)found that α for CpG transitions depends on the chromosome comparison used for estimation. Comparisons involving the Y chromosome generally suggested that CpG transitions have a weaker than average male bias, but $\frac{X}{A}$ comparisons indicated the opposite (Smith and Hurst, 1999). Within a chromosomal class, the effect of CpG sites on male-bias estimates was locus-specific. Other studies, including the most recent and comprehensive study by Taylor et al. (2006), have found that CpG transitions show less male-bias than non-CpG substitutions for $\frac{Y}{X}$ (Erlandsson et al., 2000) and $\frac{X}{A}$ (Taylor et al., 2006, Nachman and Crowell, 2000) comparisons.

Methodological biases and region-specific substitution rates may have contributed to the considerable variation in previous estimates of male-bias for CpG sites. The assignment of aligned sites into CpG and non-CpG classes is usually done parsimoniously. Because of the very high substitution rate of CpG dinucleotides compared to other types of substitutions, the use of parsimony to assign ancestral states can result in substantial bias even for closely related sequences such as human and chimpanzee (Gaffney and Keightley, 2008, Hernandez, Williamson and Bustamante, 2007), thus a likelihood approach as used in this study is preferable. Another potential source of bias is that the models used to estimate the substitution rate of CpG and non-CpG sites typically assume the substitution process is stationary and reversible. Applying this process to the subset of observed $CG \rightarrow TG$ transitions clearly violates the model assumptions. This study examined male-bias using a model designed to incorporate the

CpG substitution process into the complete substitution spectrum, and a large, genome-wide data set to facilitate detection of locus-specific effects.

This analysis first reconsiders Taylor et al.'s (2006) finding that CpG substitutions are less male-biased than non-CpG substitutions, by estimating male-bias for all dinucleotide substitutions in intergenic and intronic alignments. The question of whether instantaneous substitution rate estimates support a difference in the intrinsic mutability of CpG dinucleotides in the male and female germlines is addressed, and the relative contributions of motif frequency and transition rate differences to the male-bias for CpG substitutions examined.

4.2 Methods

A set of 24 models, referred to as the complementary dinucleotide model set and consisting of the CpG baseline model plus the 23 models containing the CpG baseline parameters and an additional strand complementary dinucleotide parameter pair, was used in this chapter. Sex-bias was estimated for each instantaneous dinucleotide substitution by fitting each model in the complementary dinucleotide model set to the intronic and intergenic alignments. Miyata et al.'s (1987) α statistic was used to estimate the strength of male-bias in the substitution process, using the same procedure as for estimating bias in the substitution rate. Estimates of α were calculated using the ratio of the mean parameter estimates for sets of X-linked and autosomal data. When estimating male-bias in the substitution process, the Q matrix was scaled so that the GTR parameter estimates sum to one (see Chapter 2, Section 2.1.6). Note that the scaling of the Q matrix also constrains estimates of α for the substitution process. The GTR parameters cannot all be greater in autosomal than in X-linked data and also sum to one. Bootstrap confidence intervals for α statistics were estimated as described in Chapter 3.

Estimates of α for dinucleotide substitution rates are presented in three ways: as the bias for the single-nucleotide substitution class of which the dinucleotide substitution is a member, e.g. $CG \leftrightarrow TG$ is an instance of a $C \leftrightarrow T$ substitution; as the bias which can be attributed to the context, e.g. how is the bias for $CG \leftrightarrow TG$ substitutions different from that of $C \leftrightarrow T$ substitutions in general; and the total bias for the dinucleotide substitutions. The latter is calculated using the product of the relevant GTR and dinucleotide terms in the model.

4.2.1 Expected number of CpG substitutions per site

The method used to estimate the expected number of CpG substitutions per site was an extension of that developed by Goldman and Yang (1994) and first applied to estimate the expected number of synonymous and non-synonymous substitutions per site for coding data. For a given branch, the Q matrix was scaled so that the branch length estimate was equal to one. The expected number of substitutions of a given type is then the sum of the relevant Q matrix entries each multiplied by the frequency of their starting motifs. Parameter values for this analysis were estimated using a model with rate parameters for transitions, transversions, CpG transitions and CpG transversions. Branch lengths were scaled to reflect the independent contributions of each of these terms. Note that the expected number of substitutions of a given type per site does not have a clear biological meaning (see discussion in Muse and Gaut, 1994, Muse, 1996)

4.3 Results

4.3.1 CpG transitions are not weakly male-biased

For each instantaneous, dinucleotide substitution, α was estimated for the total substitution rate, and the single nucleotide and dinucleotide substitution rate components that together compose the total rate. Estimates of average, genome-wide $\hat{\alpha}$ for the integenic data set are shown in Table 4.1. The $\hat{\alpha}$ statistic is used here to rank dinucleotide contexts from least to most male-biased. Whilst the $r_{i,j}$ (instantaneous substitution rate) terms for some parameters indicate a female-bias, i.e. $\hat{\alpha} < 1$, when substitution frequencies are considered all substitutions were male-biased because the autosomal alignments evolve at a faster rate than the X-linked alignments. Consequently, α estimates are described as indicating either a strong or weak male bias, rather than a male or a female bias.

Genome-wide estimates of $\hat{\alpha}$ indicate that the CpG transitions are neither weakly nor strongly biased in comparison with other types of substitutions. Estimates for the two CpG transition parameters indicate a reduction in the average male-bias for transitions at CpG dinucleotides compared with the respective GTR substitution rates, i.e., the $\hat{\alpha}_{dinuc}$ entries in Table 4.1 are less than one. This was in part due to the fact that in other sequence contexts, transitions showed the strongest male-bias. Of the eight most male-biased dinucleotide substitutions, seven were transitions. However, CpG transitions were not the least biased of the transition rates, and comparison of the confidence intervals indicates that the total bias estimated for CpG transitions does not differ from the bias estimated for the majority of the other dinucleotide contexts. In contrast, transversions involving CpG dinucleotides showed the weakest male bias, accounting for four of the five lowest ranks.

No significant difference was observed between estimates of $\hat{\alpha}$ for CpG transitions calculated using different substitution model parameterisations. Estimates of $\hat{\alpha}_{T\leftrightarrow C \times TG\leftrightarrow CG}$ and $\hat{\alpha}_{A\leftrightarrow G \times CA\leftrightarrow CG}$ were calculated for each of the complementary dinucleotide models. In all cases the estimates fell within the confidence interval estimated for the CpG baseline model and shown in Table 4.1.

Table 4.1: Total, context-dependent and context-independent $\hat{\alpha}$ estimates for dinucleotide substitution rates in intergenic data. Estimates were calculated using the average substitution rate estimates from autosomal and X-linked intergenic data for GTR parameters ($\hat{\alpha}_{GTR}$), dinucleotide parameters ($\hat{\alpha}_{dinuc}$) and their product ($\hat{\alpha}_{GTR\times dinuc}$). Estimates in brackets the 95% bootstrap confidence intervals. Parameters are ranked from most to least male-biased. Abbreviations: **Ts** - transition, **Tv** - transversion, **CpG** - substitution involving a CpG dinucleotide. Values are shown to 2 decimal places.

Rank	Context	GTR	$\hat{\alpha}_{GTR\times dinuc}$	$\hat{\alpha}_{GTR}$	$\hat{\alpha}_{dinuc}$	Type
1	$AT \leftrightarrow GT$	$A \leftrightarrow G$	1.55 (1.40–1.71)	1.07 (1.00–1.13)	1.45 (1.34–1.57)	Ts
2	$AT \leftrightarrow AC$	$T \leftrightarrow C$	1.38 (1.25–1.53)	1.02 (0.96–1.08)	1.36 (1.25–1.46)	Ts
3	$TA \leftrightarrow TG$	$A \leftrightarrow G$	1.30 (1.18–1.42)	1.09 (1.02–1.15)	1.19 (1.11–1.28)	Ts
4	$TA \leftrightarrow CA$	$T \leftrightarrow C$	1.24 (1.13–1.37)	1.02 (0.96–1.08)	1.22 (1.12–1.32)	Ts
5	$CC \leftrightarrow GC$	$C \leftrightarrow G$	1.14 (0.96–1.36)	1.01 (0.94–1.08)	1.13 (0.93–1.35)	Tv
6	$AG \leftrightarrow GG$	$A \leftrightarrow G$	1.12 (1.01–1.23)	1.10 (1.02–1.17)	1.01 (0.91–1.12)	Ts
7	$CT \leftrightarrow CC$	$T \leftrightarrow C$	1.11 (1.02–1.22)	1.02 (0.96–1.09)	1.08 (0.98–1.18)	Ts
8	$AC \leftrightarrow GC$	$A \leftrightarrow G$	1.10 (0.98–1.23)	1.09 (1.02–1.16)	1.01 (0.93–1.11)	Ts
9	$CA \leftrightarrow GA$	$C \leftrightarrow G$	1.06 (0.91–1.24)	0.98 (0.91–1.06)	1.08 (0.93–1.27)	Tv
10	$TG \leftrightarrow AG$	$T \leftrightarrow A$	1.06 (0.88–1.24)	0.74 (0.68–0.80)	1.44 (1.23–1.71)	Tv
11	$TC \leftrightarrow TG$	$C \leftrightarrow G$	1.05 (0.91–1.21)	0.98 (0.91–1.06)	1.08 (0.93–1.24)	Tv
12	$TT \leftrightarrow TC$	$T \leftrightarrow C$	1.03 (0.94–1.13)	1.02 (0.96–1.09)	0.99 (0.90–1.10)	Ts
13	$AA \leftrightarrow GA$	$A \leftrightarrow G$	1.01 (0.92–1.13)	1.10 (1.03–1.18)	0.91 (0.83–1.01)	Ts
14	$CC \leftrightarrow CA$	$C \leftrightarrow A$	1.01 (0.89–1.18)	0.89 (0.84–0.95)	1.13 (0.99–1.34)	Tv
15	$TG \leftrightarrow GG$	$T \leftrightarrow G$	0.98 (0.84–1.14)	0.84 (0.79–0.90)	1.16 (1.00–1.34)	Tv
16	$AC \leftrightarrow AG$	$C \leftrightarrow G$	0.97 (0.84–1.13)	1.03 (0.96–1.11)	0.93 (0.81–1.07)	Tv
17	$CC \leftrightarrow AC$	$C \leftrightarrow A$	0.97 (0.83–1.13)	0.90 (0.84–0.96)	1.09 (0.94–1.26)	Tv
18	$GC \leftrightarrow GG$	$C \leftrightarrow G$	0.96 (0.81–1.13)	1.01 (0.93–1.08)	0.95 (0.80–1.14)	Tv
19	$GA \leftrightarrow GG$	$A \leftrightarrow G$	0.95 (0.86–1.06)	1.10 (1.03–1.17)	0.86 (0.78–0.95)	Ts
20	$GT \leftrightarrow GC$	$T \leftrightarrow C$	0.95 (0.85–1.05)	1.03 (0.98–1.10)	0.92 (0.84–1.01)	Ts
21	$TC \leftrightarrow TA$	$C \leftrightarrow A$	0.95 (0.81–1.10)	0.90 (0.84–0.97)	1.05 (0.91–1.22)	Tv
22	$TG \leftrightarrow CG$	$T \leftrightarrow C$	0.94 (0.82–1.08)	1.03 (0.97–1.09)	0.91 (0.80–1.04)	CpG Ts
23	$CT \leftrightarrow GT$	$C \leftrightarrow G$	0.94 (0.81–1.09)	1.03 (0.96–1.12)	0.89 (0.76–1.04)	Tv
24	$CA \leftrightarrow CG$	$A \leftrightarrow G$	0.91 (0.77–1.06)	1.09 (1.02–1.17)	0.83 (0.72–0.96)	CpG Ts
25	$GC \leftrightarrow GA$	$C \leftrightarrow A$	0.90 (0.75–1.08)	0.91 (0.85–0.97)	0.99 (0.82–1.17)	Tv
26	$TA \leftrightarrow AA$	$T \leftrightarrow A$	0.86 (0.74–1.00)	0.77 (0.70–0.84)	1.10 (0.96–1.25)	Tv
27	$CT \leftrightarrow AT$	$C \leftrightarrow A$	0.86 (0.74–0.99)	0.91 (0.85–0.97)	0.94 (0.81–1.10)	Tv
28	$TT \leftrightarrow GT$	$T \leftrightarrow G$	0.86 (0.75–0.98)	0.85 (0.80–0.92)	1.00 (0.88–1.13)	Tv
29	$TC \leftrightarrow CC$	$T \leftrightarrow C$	0.85 (0.76–0.94)	1.04 (0.98–1.10)	0.82 (0.74–0.90)	Ts
30	$GT \leftrightarrow GG$	$T \leftrightarrow G$	0.85 (0.71–1.00)	0.86 (0.80–0.92)	0.98 (0.83–1.17)	Tv
31	$CT \leftrightarrow CA$	$T \leftrightarrow A$	0.84 (0.71–1.00)	0.74 (0.67–0.80)	1.14 (0.96–1.35)	Tv
32	$TC \leftrightarrow GC$	$T \leftrightarrow G$	0.84 (0.70–0.99)	0.86 (0.80–0.92)	0.97 (0.80–1.15)	Tv
33	$AC \leftrightarrow AA$	$C \leftrightarrow A$	0.83 (0.72–0.95)	0.92 (0.86–0.98)	0.90 (0.78–1.04)	Tv
34	$TT \leftrightarrow CT$	$T \leftrightarrow C$	0.82 (0.75–0.90)	1.04 (0.97–1.10)	0.78 (0.71–0.85)	Ts
35	$TT \leftrightarrow TG$	$T \leftrightarrow G$	0.81 (0.71–0.93)	0.86 (0.81–0.92)	0.94 (0.83–1.08)	Tv
36	$AA \leftrightarrow AG$	$A \leftrightarrow G$	0.81 (0.73–0.88)	1.12 (1.04–1.19)	0.72 (0.66–0.78)	Ts
37	$CA \leftrightarrow AA$	$C \leftrightarrow A$	0.80 (0.71–0.90)	0.92 (0.86–0.99)	0.86 (0.76–0.98)	Tv
38	$AT \leftrightarrow AG$	$T \leftrightarrow G$	0.80 (0.69–0.92)	0.86 (0.81–0.93)	0.92 (0.80–1.05)	Tv
39	$TA \leftrightarrow GA$	$T \leftrightarrow G$	0.79 (0.68–0.91)	0.87 (0.81–0.93)	0.91 (0.79–1.05)	Tv
40	$TC \leftrightarrow AC$	$T \leftrightarrow A$	0.76 (0.62–0.93)	0.77 (0.71–0.84)	0.97 (0.81–1.18)	Tv
41	$TT \leftrightarrow AT$	$T \leftrightarrow A$	0.75 (0.63–0.88)	0.81 (0.74–0.88)	0.94 (0.80–1.08)	Tv
42	$TT \leftrightarrow TA$	$T \leftrightarrow A$	0.74 (0.64–0.87)	0.77 (0.70–0.85)	0.93 (0.79–1.11)	Tv
43	$GT \leftrightarrow GA$	$T \leftrightarrow A$	0.74 (0.59–0.90)	0.77 (0.71–0.84)	0.95 (0.78–1.16)	Tv
44	$CG \leftrightarrow AG$	$C \leftrightarrow A$	0.73 (0.58–0.90)	0.89 (0.83–0.95)	0.82 (0.65–1.02)	CpG Tv
45	$CG \leftrightarrow GG$	$C \leftrightarrow G$	0.68 (0.51–0.91)	1.02 (0.94–1.10)	0.64 (0.47–0.88)	CpG Tv
46	$AT \leftrightarrow AA$	$T \leftrightarrow A$	0.63 (0.54–0.74)	0.81 (0.74–0.88)	0.79 (0.68–0.92)	Tv
47	$CC \leftrightarrow CG$	$C \leftrightarrow G$	0.61 (0.46–0.81)	1.02 (0.94–1.09)	0.57 (0.41–0.75)	CpG Tv
48	$CT \leftrightarrow CG$	$T \leftrightarrow G$	0.61 (0.46–0.78)	0.84 (0.78–0.91)	0.72 (0.55–0.92)	CpG Tv

4.3.2 Weak bias for CpG transversions is not genome-wide.

The weak genome-wide male-bias observed for CpG transversions in Table 4.1 is a consequence of differences in CpG transversion rates amongst autosomal alignments rather than a difference in CpG mutability between the male and female germlines. Wilcoxon Signed-rank tests were performed to determine whether dinucleotide rate estimates differed between the X-chromosome and any of the autosomes. A conservative Bonferroni correction was used to adjust the p-value of 0.05 to account for the 22 comparisons performed for each of the 48 dinucleotide parameters. The significant comparisons are shown in Table 4.2. CpG transition and transversion rate estimates only differed between the X chromosome and the smaller-sized autosomes. Chromosome-specific relative rate estimates for a CpG transition and transversion parameter, and the most biased dinucleotide parameter are shown in Supplementary Material, Figure 7.1. Previous analyses (Chapter 3, Section 3.3.2) indicated that the dinucleotide motif composition of the intergenic, X-linked data is most similar to that of chromosomes 3, 5 and 6. Estimating α as in Table 4.1 by comparing mean rate estimates from these autosomes with X-linked estimates eliminated the apparent weak bias for CpG transversions (Supplementary Material, Table 7.1). Bias estimates for other types of substitutions were largely unaffected.

4.3.3 Influence of methylation in intronic data

Estimates of $\hat{r}_{TG\leftrightarrow CG}$ and $\hat{r}_{CA\leftrightarrow CG}$ tended to be lower in intronic than intergenic data (Figure 4.1). The rate reduction was more pronounced for $\hat{r}_{CA\leftrightarrow CG}$ than for $\hat{r}_{TG\leftrightarrow CG}$ estimates. This result may indicate that the rate of $CA \leftrightarrow CG$ substitution is reduced on the non-transcribed strand of genes, or equivalently

Table 4.2: Differences between dinucleotide substitution rate estimates for intergenic X-linked and autosomal alignments. For each parameter, substitution rate estimates from X-linked alignments were compared with estimates from each of the autosomes using a Wilcoxon signed-rank test. Comparisons that were significant at the 0.05 level after application of a Bonferroni multiple test correction are shown. Dinucleotide parameters are sorted according to the number of significant comparisons between the X chromosome and the autosomes. Estimates of the parameters not listed did not differ between the X chromosome and any of the autosomes.

Context	Substitution	Significant comparisons
$AA \leftrightarrow AG$	$A \leftrightarrow G$	1–20, 22
$AT \leftrightarrow GT$	$A \leftrightarrow G$	1–18, 20, 22
$AT \leftrightarrow AC$	$T \leftrightarrow C$	1–12, 14–21
$TT \leftrightarrow CT$	$T \leftrightarrow C$	2, 6–12, 14–17, 20
$TC \leftrightarrow CC$	$T \leftrightarrow C$	7, 9–11, 15–18, 20–22
$TA \leftrightarrow CA$	$T \leftrightarrow C$	2, 3, 5, 7, 9, 11, 12, 15
$TA \leftrightarrow TG$	$A \leftrightarrow G$	2, 9–11, 15, 16
$CT \leftrightarrow CG$	$T \leftrightarrow G$	15–17, 19, 22
$CA \leftrightarrow CG$	$A \leftrightarrow G$	16, 17, 19, 22
$TG \leftrightarrow CG$	$T \leftrightarrow C$	16, 17, 19, 22
$CG \leftrightarrow AG$	$C \leftrightarrow A$	10, 16, 17, 22
$CC \leftrightarrow CG$	$C \leftrightarrow G$	10, 15, 16, 21
$CG \leftrightarrow GG$	$C \leftrightarrow G$	10, 15, 18, 22
$GA \leftrightarrow GG$	$A \leftrightarrow G$	10, 16, 17, 22
$TG \leftrightarrow AG$	$T \leftrightarrow A$	4,13
$AC \leftrightarrow GC$	$A \leftrightarrow G$	20

that the rate of $TG \leftrightarrow CG$ substitution is reduced on the transcribed strand. Estimates of $\hat{\alpha}$ for the intronic data set are shown in Supplementary Material, Table 7.2. Although $\hat{\alpha}$ statistics for the CpG transversions suggested that male-bias is weaker in the intronic than the intergenic data set, the 95% confidence intervals for these statistics overlapped indicating that the difference was not significant. As in intergenic data, genome-wide $\hat{\alpha}$ estimates from introns indicated that CpG transversions are the least male-biased. It is assumed but not tested here that this result again reflects compositional differences between the autosomal and X-linked data sets.

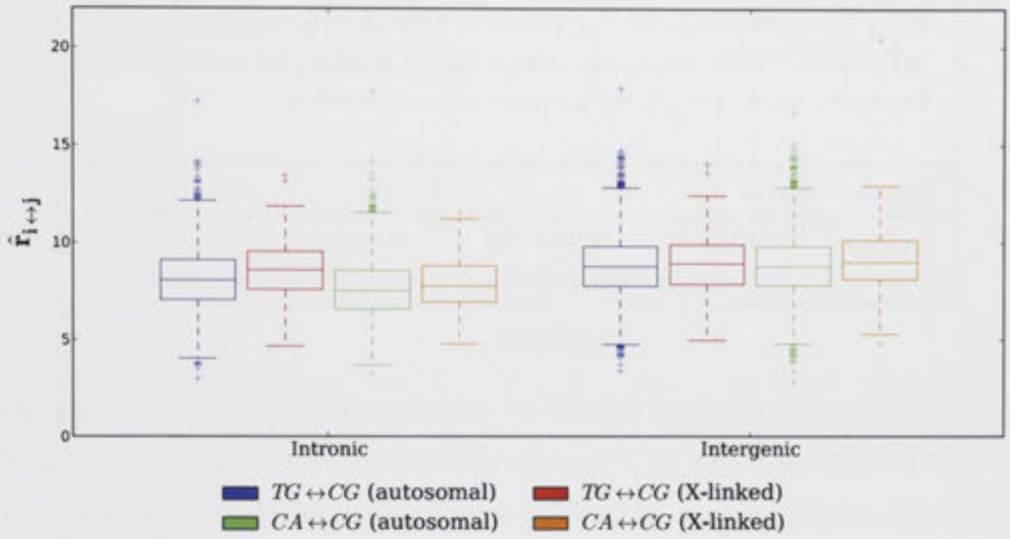

Figure 4.1: CpG transition rate estimates for intronic and intergenic data. Estimates for the CpG baseline model are shown for data from the entire autosomal and the X-linked data sets. Whiskers and outliers are defined as in Figure 3.1.

4.3.4 Effect of motif frequency differences on male-bias estimates

Whilst it was shown in Table 4.1 that the instantaneous CpG transition rate is approximately equal in X-linked and autosomal data, the CpG substitution frequency along the human branch exhibits an above-average male bias as a consequence of the often greater CpG dinucleotide frequency in autosomal than in X-linked alignments (see Chapter 2, Figure 3.1). Figure 4.2 shows the number of transitions, CpG transitions, transversions and CpG transversions expected along the human branch, and $\hat{\alpha}$ statistics calculated by comparing the mean number of substitutions of each type are shown in Table 4.3. Considering the rarity of CpG dinucleotides, CpG transitions account for a large fraction of the substitution rate difference between X-linked and autosomal alignments.

Table 4.3: **$\hat{\alpha}$ estimates for CpG and non-CpG transitions and tranversions.** Estimates were calculated by comparing the average expected number of substitutions in X-linked and autosomal data. Abbreviations **CpG ts** transition involving a CpG, **ts** transition not involving a CpG, **CpG tv** transversion involving a CpG, **tv** transversion not involving a CpG

Data type	CpG ts	ts	CpG tv	tv
Intergenic	53.63 (20.06–∞)	6.48 (5.01–8.52)	13.14 (8.25–27.93)	4.32 (3.36–5.76)
Intronic	∞ (∞–)	234.32 (18.45–∞)	∞ (∞–)	24.89 (9.24–∞)

Based on the mean expected number of substitutions per site, CpG transitions account for approximately 14-15% of the substitution rate difference between X-linked and autosomal data, and CpG transversions another 3%. CpG transition frequencies exhibit a stronger male-bias than all of the other substitution types considered, and CpG transversion frequencies exhibit a stronger male-bias than non-CpG transversions.

4.4 Discussion

Taylor et al. (2006) found that substitutions at CpG sites show less male-bias than substitutions at non-CpG sites, and concluded that that CpG and non-CpG substitutions occur via different mechanisms. This study expanded on these results by considering male-bias for CpG substitutions in relation to the complete dinucleotide substitution spectrum. It was supposed that if CpG transitions have a weak male-bias because they occur via a replication-independent mechanism

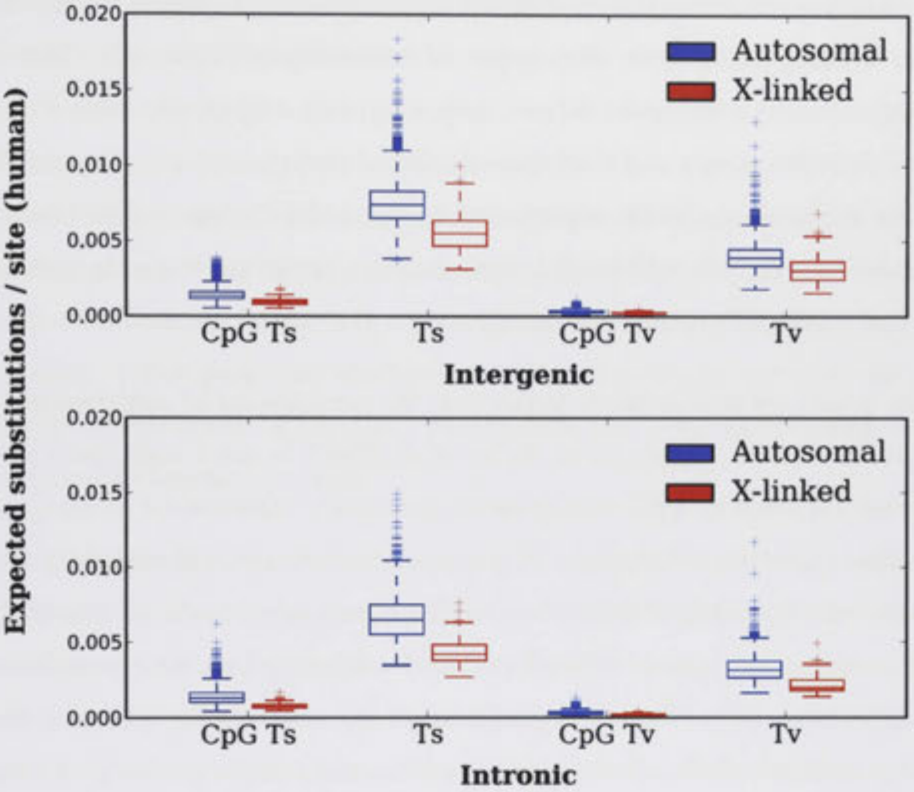

Figure 4.2: Contributions of transitions and transversions within and outside the CpG context to the human substitution rate. Substitution frequencies are estimated as the number of substitutions per dinucleotide site expected along the human branch of the alignments. Abbreviations **CpG ts** transition involving a CpG, **ts** transition not involving a CpG, **CpG tv** transversion involving a CpG, **tv** transversion not involving a CpG

and all other substitutions occur via a replication-dependent mechanism, CpG transitions should be less biased than all other types of substitutions.

The results presented in this chapter are in conflict with the results presented by Taylor et al. (2006). By considering genome-wide average substitution rate estimates, it was observed here that the CpG transversion rate is weakly male-biased in comparison with other types of substitutions (Table 4.1). However, this result was not observed when comparing only alignments with a similar dinucleotide frequency and G+C% composition (Supplementary Material Table 7.1). As Taylor et al. (2006) employed a correction for G+C% in their study, the results of this study would predict that no difference in the the male bias in the CpG and non-CpG substitution rate should be observed.

Results presented in this study imply that the partitioning of substitutions into CpG and non-CpG classes, as in Taylor et al. (2006), is not a sufficient measure of evaluating whether CpG substitutions are weakly male-biased in comparison with other types of substitutions. If genome-wide estimates of male-bias for the CpG and non-CpG substitution rates had been compared, the weak bias for CpG transversions might result in the CpG class exhibiting less male-bias than the non-CpG class. However, other partitions of the substitution spectrum would produce similar results. For example, substitutions between T and A show a weaker than average male-bias (see Table 4.1). If substitutions were partitioned into $T \leftrightarrow A$ transversions and non- $T \leftrightarrow A$ substitutions, $T \leftrightarrow A$ transversions would show a weak male-bias. The results of this study therefore do not support the hypothesis that CpG substitutions occur independently of replication and exhibit a weak male-bias in comparison with all non-CpG substitutions.

Implications for understanding for the mechanism of CpG mutation

The observed male-bias for CpG transitions was predicted based on methylation pattern differences between the male and the female germlines. However, it was not expected that the strength of the male-bias for CpG substitutions should equal that observed for other types of substitutions, as CpG lesions arise via a unique mechanism. One simple explanation for this observation is that male bias results from differences in repair activity between the male and female germlines, rather than differences in the number of replication cycles. As nucleotide excision repair (NER) and BER target a wide range of lesions, sex differences in the activity or expression of any of the core components of these repair systems could result in a sex bias of consistent magnitude across the entire substitution spectrum. The implication of this proposed mechanism is that the rate of repair is the most important determinant of the CpG substitution rate after methylation status, rather than the deamination rate.

Even supposing that the replication rate is the major determinant of the CpG substitution rate, the consistent strength of male-bias for CpG and other types of substitutions is not easily explained by the replication origin hypothesis. If the elevated CpG transition rate is the result of the cumulative effects of replication-depenendent and deamination induced substitutions, there is no reason to suppose that the male-bias for CpG transitions should be equal to the bias for a substitution type caused solely by replication.

Contribution of methylation to male-bias

The frequency of CpG transitions on the human lineage in X-linked and autosomal alignments showed a stronger male-bias than the frequency of non-CpG transitions or transversions (Table 4.3). As the rate of CpG transitions tended to be less-biased than the rate of other transitions, this result is primarily a consequence of the elevated CpG transition rate in combination with CpG frequency differences between the X chromosome and the majority of the autosomes. Krawczak et al. (1998) similarly found that CpG transitions constitute a larger proportion of autosomal than X-linked mutations associated with human disease because of differences in the frequency of CpG dinucleotides, and unusually low values of α have been estimated for sequences with a low CpG frequency (McVean, 2000, Bohossian et al., 2000). The X chromosome and the autosomes may differ in the frequency of CpG dinucleotides as a result of the repetitive elements they contain. The human X chromosome contains proportionately more LINE1 elements and less Alu elements than the autosomes (Smit, 1999). Alus have a high CpG frequency whilst LINE1s have a low frequency (Ohshima et al., 2003). It has been proposed that the unique repeat distribution of the X chromosome is involved in X inactivation (Chow et al., 2005). Motif frequency differences may contribute to differences between species in α estimates. The difference between the CpG frequencies of X-linked and autosomal sequences is more pronounced in rodents than in humans (Jensen-Seaman et al., 2004), suggesting that CpG substitutions may account for a greater proportion of the male bias in rodents.

If the elevated α estimates for CpG transitions are a consequence of motif frequency differences, these estimates derive from an effect of chromosomal location rather than sex. No direct effect of methylation on male-bias was detected. However the fact that the strength of the male-bias in the CpG transition rate was similar to strength of the bias for most other types of substitution may indicate that methylation has an indirect effect on male-bias. Methylation is closely associated with chromatin condensation (e.g. Stancheva, 2005), gene expression (Fuks, 2005) and recombination rates (Sigurdsson et al., 2009), all of which are known to affect mutation. Methylation may contribute to the male-bias by differentially marking sequences in the male and the female germlines, thus altering how they are affected by other agents. It remains to be determined whether CpG substitutions exhibit an average male bias because methylation indirectly affects male bias, or because the processes that cause male-bias affect CpG and non-CpG dinucleotides equivalently. However, any effect of methylation on the male-bias does not appear to be a simple consequence of methylation differences between the male and female germlines.

Chapter 5

Transcription

In this chapter, the contribution of transcription to male bias estimates is evaluated for the human, chimpanzee and macaque lineages. I consider whether the effects of transcription are modulated by the background substitution process, and whether there is a simple relationship between strand asymmetry and male bias in the substitution process.

5.1 Introduction

Transcription is a plausible contributor to male-biased evolution because transcription affects the nucleotide substitution process (Green et al., 2003), and sex-specific programs of transcription play important roles in gametogenesis. Like methylation, transcription has a predictable effect on the substitution spectrum, which allows its contribution to male-biased evolution to be evaluated. The effect of transcription is measured as a difference in the rates of strand complementary substitutions, i.e. asymmetric substitution rates.Intronic sequences from human and chimpanzee are distinguished from neighbouring intergenic sequences by a greater difference between the substitution processes of complementary DNA strands, measured as a difference in the rates of strand complementary substitutions (Green et al., 2003). As similar substitution rate asymmetries, believed to result from replication, are seen in intergenic sequences (Touchon et al., 2005), the effects of transcription should be measured relative to the background rate for a given region.

5.1.1 Effects of transcription on substitution rate and process

Transcription increases strand asymmetry in the substitution process because it affects the two strands of a DNA molecule differently. The non-transcribed (template) strand is left in a single-stranded state, leaving it vulnerable to damage, whilst the transcribed (non-template) strand is transiently paired with the nascent RNA and occluded by the RNA polymerase. Additionally, a repair system known as transcription-coupled repair (TCR) specifically repairs only the transcribed strand. TCR is a subpathway of NER, which repairs a range of bulky DNA lesions (see Hanawalt and Spivak, 2008, for a review of TCR). The other component of NER is global genomic repair (GGR), which can repair both strands of a transcribed gene as well as the surrounding region. In the absence of TCR, the RNA polymerase can hinder NER, probably by physically blocking access to damaged DNA (Li and Smerdon, 2004).

Though asymmetry is a convenient indicator of transcription, transcription can also affect the mutation rate of the region surrounding a gene without necessarily increasing strand asymmetry. Transcription is associated with changes in chromatin structure that are conducive both to binding of transcription factors and to DNA repair (Workman and Kingston, 1998). Open chromatin is thought to passively facilitate repair by making the DNA more accessible. Transcription activators can also actively promote repair independently of TCR, by recruiting repair enzymes to transcription initiation sites (Frit et al., 2002). Inverse associations have been detected between histone density and transcription intensity in yeast and *Drosophila* (in Williams and Tyler, 2007); and aligned primate sequences with an open chromatin structure in the modern human genome show less divergence than those with a condensed structure (Ying et al., 2010, Prendergast et al., 2007). These results suggest that transcription should decrease the evolutionary rate of genes and their surrounding regions.

The overall effect of transcription on the mutation rate is a combination of its repair promoting activity, mediated via chromatin structure and TCR, and mutagenic effects resulting from increased exposure of the non-transcribed strand to damage during transcription. The relative contribution of each of these factors is likely to be influenced by the surrounding region. Transcription-associated chromatin decondensation, for example, may have a negligible effect on the mutation process in a region already located in open chromatin. Transcription intensity also appears to affect transcription-associated mutagenesis (Majewski, 2003), probably because the efficiency of TCR depends on transcription intensity (Leadon and Lawrence, 1991).

Estimates of the net effect of transcription on mutation in eukaryotes have been contradictory. Reversion assays in *Saccharymyces cerevisiae* indicate that high rates of transcription are associated with an increased mutation rate (Datta and Jinks-Robertson, 1995, Kim et al., 2007), with the magnitude of the increase directly related to the intensity of expression (Kim et al., 2007). Hendriks et al. (2008)

found a similar increase in mutation in a reporter gene construct in mouse embryonic stem cells following UV treatment. In Hendriks et al.'s (2008) study, transcription status did not affect the induction of lesions, but increased the frequency of mutation in a UV dose-dependent manner. In contrast, other evidence suggests that transcription results in either a reduction (Lippert et al., 1998, Smith et al., 2002) or no change (Green et al., 2003) in mutation rate in comparison with surrounding non-transcribed regions. This may be because the mutagenic effects of transcription are balanced by TCR and transcription-associated chromatin changes that are conducive to repair. Estimates of the effect of transcription on mutation in bacteria have been similarly inconsistent. Experimental evidence tends to suggest that transcription increases the mutation rate in bacteria whilst comparative evidence tends to indicate the opposite (Ochman, 2003).

Strand asymmetry in the substitution process and in sequence composition have been used in previous studies to estimate the influence of transcription on the evolutionary process. If nucleotide substitutions occur with equal frequency on both DNA strands, within a single strand the frequency of A should equal the frequency of T, and the frequency of G should equal that of C. Strand asymmetric substitution results in deviations from this expectation, which is known as Chargaff's strand parity rule (see Sueoka (1995) and Baisnee et al. (2002)). The majority of human genes examined by Green et al. (2003) and Majewski (2003) showed an excess of G and T over A and C nucleotides on the non-transcribed (i.e. coding) strand. The extent of strand asymmetry in human housekeeping genes is correlated with their average expression intensity across a range of somatic tissues (Majewski, 2003). Both TCR and GGR efficiency also vary positively with transcription intensity and inversely with chromatin compaction (Feng, Drost, Scaringe, Liu and Sommer, 2002, Feng, Hu, Komissarova, Pao, Hung, Adair and Tang, 2002, Episkopou et al., 2009).

5.1.2 Transcription in the germline and male-bias

Chromatin structure and transcription intensity have been implicated as factors that can influence the transcription-associated mutation rate, and may contribute to male-bias if there are sex-differences in the interplay between transcription, repair and mutation. For example, assuming transcription is protective, more intensive transcription in the female than the male germline would lead to a male-biased substitution rate and process. Notable stages of sex-specific germline transcription programs are outlined below. The effect of transcription on male-bias is difficult to predict, as gene expression status and intensity varies between genes and between stages of gametogenesis.

Transcription levels fluctuate throughout oogenesis. Gene expression is very intensive and widespread throughout the oocyte growth period, and decreases dramatically once growth is completed before the end of meiosis I (Picton et al., 1998). During the intense growth period, oocytes store mRNA to support the embryo after fertilisation until the embryonic genome is activated (reviewed in Picton et al., 1998). The oocyte growth phase is probably the most intensive period of widespread transcription in either germline, but is of a short duration compared to other stages of oogenesis. Oocytes spend a long time period, up to decades in humans, in a state of quiescence. Primary follicles are formed prenatally and undergo little subsequent transcription until puberty (Eichenlaub-Ritter and Peschke, 2002).

Spermatogenesis is associated with changes in chromatin structure and gene expression. In the post-meiotic stage of spermatogenesis, the chromatin structure is dramatically changed and most of the cytoplasm is lost. A unique feature of spermatogenesis is that transcription occurs in the haploid spermatid autosomes until the replacement of histones with protamines (reviewed in Dadoune et al., 2004). Although RNA is present in mature spermatozoa, there is no evidence of active transcription in the nuclear genome (Grunewald et al., 2005). Peaks in transcriptional activity occur in the mitotic phase, the start of meiosis and immediately preceding chromatin condensation in the spermatid (Wrobel and Primig, 2005). Most of the spermatogenesis-specific transcription occurs either at or post-meiosis (Schultz et al., 2003).

Repair activity declines during the late stages of spermatogenesis. For certain types of lesions, intact rat meiotic and post-meiotic male germ cells show inefficient or non-existant NER activity despite earlier spermatogenic cell stages showing high activity (Jansen et al., 2001). Another study found that different types of lesions were repaired efficiently by NER in various mouse spermatogenic cell stages (Xu et al., 2005), although in both of these studies repair was less efficient than in somatic cells. NER activity in postmeiotic mouse spermatids was significantly reduced in middle-aged compared with young animals (Xu et al., 2005). TCR of lesions resulting from UV exposure was dependent on developmental stage in mouse male germ cells, with repair in mitotic cell stages more efficient than repair in meiotic or post-meiotic cells. However, inefficient NER in rodent male germ cells does not imply that damage is transmitted. Xu et al. (2005) proposed that damaged spermatogenic cells may be targeted for apoptosis rather than repaired.

The distribution of genes according to their function and timing of expression is non-random (Khil et al., 2004), which could affect α estimates. In the male germline, sex chromosomes are inactivated at meiosis and most sex-linked genes remain transcriptionally repressed for the post-meiotic period (Turner, 2007). If transcription is mutagenic, the sex chromosomes may be less affected by transcription than the autosomes, as they are generally expressed for a shorter period of time in the male germline. This could lead to a male-bias for X-autosome comparisons, and a stronger female bias for Y-autosome comparisons.

In this chapter, the contribution of transcription to the estimated male-bias in the substitution rate is evaluated. As in previous analyses, male-bias was estimated using Miyata et al.'s (1987) α statistic, by comparing the substitution rate and process of X-linked and autosomal alignments. Note that the relative time a gene spends in the male and female germlines is not necessarily related to the relative influences of male and female germline transcription on its mutation rate. Whilst X-linked genes spend more time in the female than the male germline, they may still be expressed predominantly or only in the male germline. This study only deals with the question of whether transcription affects estimates of α, due to the difficulty of estimating average transcription intensity throughout gametogenesis and the lack of sufficient gene expression data for oogenesis in humans.

5.2 Methods

Pairs of flanking alignments were selected from the intronic and intergenic data sets according to the criterion that no more than 100 000 nucleotides separated the end of one sequence from the start of its flanking pair in any species. Intergenic alignments that flanked either the start or the end of an intronic gene were included. A total of 919 flanking pairs of alignments were identified, consisting of 895 pairs of autosomal alignments and 24 pairs of X chromosomal alignments. The intronic alignments were sampled with respect to the transcribed DNA strand. The intergenic alignments were sampled from the strand corresponding to the transcribed strand of the neighbouring intronic alignment. The intergenic alignments are putatively untranscribed, or at least, transcribed less intensely than known genes. Male bias estimates and confidence intervals were calculated as described in Chapter 3,

Gene annotation data was obtained from Ensembl release 50. All human gene descriptions were searched for the keywords *testis*, *sperm*, *ovary*, *oocyte*, *oogenesis*, *egg*, *germ* and *meiosis*. The keywords were allowed to occur anywhere within the description, for example the search term *sperm* retrieved gene descriptions containing *sperm*, *spermatocyte*, *azoospermia*, etc. Of the entire Ensembl human gene collection, 434 genes were detected using the keyword search and 12 of these were represented in the flanking alignment pair data set. The descriptions for these genes were manually classified into categories: 5 *'sperm'* genes with annotations including the word *'sperm'*, 1 *'ovary'* gene defined similarly, and 6 *'testis-specific'* genes expressed specifically in the testis. The single 'ovary' gene was not given special consideration in analyses.

Multiple regression analysis

The significance of chromosomal location (autosomal or X-linked) as a predictor of substitution rate was assessed using multiple linear regression. A separate regression model was fitted for each branch, using substitution rate estimates from the flanking pair data set. Intronic substitution rates were modeled as a linear function of the substitution rate of neighbouring intergenic alignments and chromosomal location. One X-linked alignment pair with an atypically low intronic substitution rate in the macaque lineage was excluded from the macaque analysis. Normality of the residuals was verified by visual inspection of the quantile-quantile plot of theoretical versus observed residuals. Regression analysis was performed using R version 2.7.2.

5.3 Results

As the effects of transcription on the nucleotide substitution rate potentially extend beyond the transcribed region, neighbouring intronic and intergenic sequences may evolve at more similar rates than intronic and intergenic sequences on average. The contribution of transcription-related processes to the male mutation bias was thus assessed by comparing lineage-specific estimates of male bias ($\hat{\alpha}$) for neighbouring intronic and intergenic alignment pairs, and for intronic alignments in gene-rich regions versus intergenic alignments in gene-poor regions. The former comparison was made using the "flanking" data set, and the "other" remaining alignments for which no flanking alignment was available were used for the latter comparison.

When the 95% confidence intervals indicated a significant difference between lineage-specific male bias estimates ($\hat{\alpha}$) for intronic and intergenic alignment sets, the bias was consistently greater in the intronic than the intergenic set (Table 5.1). Differences in male bias were less pronounced for the "flanking" than the "other" intronic and intergenic sequences, as expected if the influence of transcription on

Table 5.1: $\hat{\alpha}$ estimates for intronic and intergenic data. α was estimated for the flanking pair data set, the complete intronic and intergenic data sets ('All Intronic' and 'All Intergenic'), and the intronic and intergenic alignments that were not members of the flanking data set ('Other Intronic' and 'Other Intergenic'). Estimates of α were calculated using the ratio of the mean X-linked and autosomal substitution rate (branch length) estimates for each data set. Branch lengths were scaled conventionally. The 95% confidence intervals are shown in brackets. Confidence intervals were calculated by resampling from each set of substitution rate estimates with replacement. The value '∞' is used when the ratio of X-linked and autosomal substitution rates ($\frac{X}{A}$) did not produce a valid α statistic, that is, $\frac{2}{3} \leq \frac{X}{A} \leq \frac{4}{3}$. Estimates are shown to 3 decimal places.

Data	**Human**	**Chimpanzee**	**Macaque**
Flanking intronic	24.741 (5.846 – ∞)	1.181 (0.721 – 1.932)	2.357 (1.300 – 3.935)
Flanking intergenic	9.147 (3.813 – 388.452)	1.302 (0.665 – 2.368)	2.116 (1.425 – 3.247)
All intronic	35.645 (9.121 – ∞)	1.624 (1.255 – 2.105)	3.072 (2.481 – 3.742)
All intergenic	6.378 (4.947 – 8.684)	1.133 (0.955 – 1.337)	1.848 (1.658 – 2.053)
Other intronic	40.059 (7.860 – ∞)	1.808 (1.344 – 2.483)	3.391 (2.838 – 4.068)
Other intergenic	6.230 (4.833 – 8.534)	1.122 (0.950 – 1.326)	1.828 (1.637– 2.040)

the substitution process extends beyond the transcribed region. Further, male bias was greatest for the "other intronic" alignments, which because of the lack of long neighbouring intronic alignments are presumably located in gene-dense regions. Similarly, male-bias was weakest for the "other intergenic" alignments, which are presumably located in gene-poor regions.

Male-bias estimates for the human and chimpanzee lineages differed significantly in each of the data sets. Estimates in Table 5.1 indicate a strong male bias for the human lineage, a weak bias for the macaque branches, and little evidence of male-bias on the chimpanzee lineage. No significant male-bias was detected on the chimpanzee branch for the intergenic data, suggesting that the bias observed in intronic data may be entirely related to transcription.

The effect of transcription on the substitution rate, measured as the difference in the substitution rate estimates of flanking intronic and intergenic sequences, was strongly related to the background, or intergenic substitution rate (Figure 5.1). In each of the species, when the autosomal, intergenic substitution rate was at the lower end of its range, the corresponding intronic rate was often faster, suggesting that transcription increases the mutation rate. This was also true of X-linked chimpanzee sequences (Figure 5.1 b). For human (Figure 5.1 a) and macaque (Figure 5.1 c) X-linked sequences however, transcription tended to decrease the mutation rate even when the background rate was low. Regression analysis confirmed that X-linked versus autosomal status was a highly significant predictor of the human intronic substitution rate even after accounting for the substitution rate of the neighbouring intergenic region ($p << 0.001$), suggesting that the increased male-bias of intronic compared to intergenic data is not a consequence of the neighbourhood in which X-linked genes are located. Chromosome class was also a nominally significant predictor of the macaque intronic substitution rate ($p < 0.01$), but not a significant predictor of the chimpanzee intronic substitution rate. The effect of transcription on autosomal testis-specific genes and genes known to be expressed in spermatogenesis, or expressed only in the male-germline was not obviously different from its effect on autosomal genes in general (Figure 5.1).

Concordance of transcription-associated male-bias with changes in the substitution process, in particular increased substitution rate asymmetry, was assessed. As expected, parameter estimates for intronic and intergenic regions were considerably different and estimates of strand complementary GTR parameters were

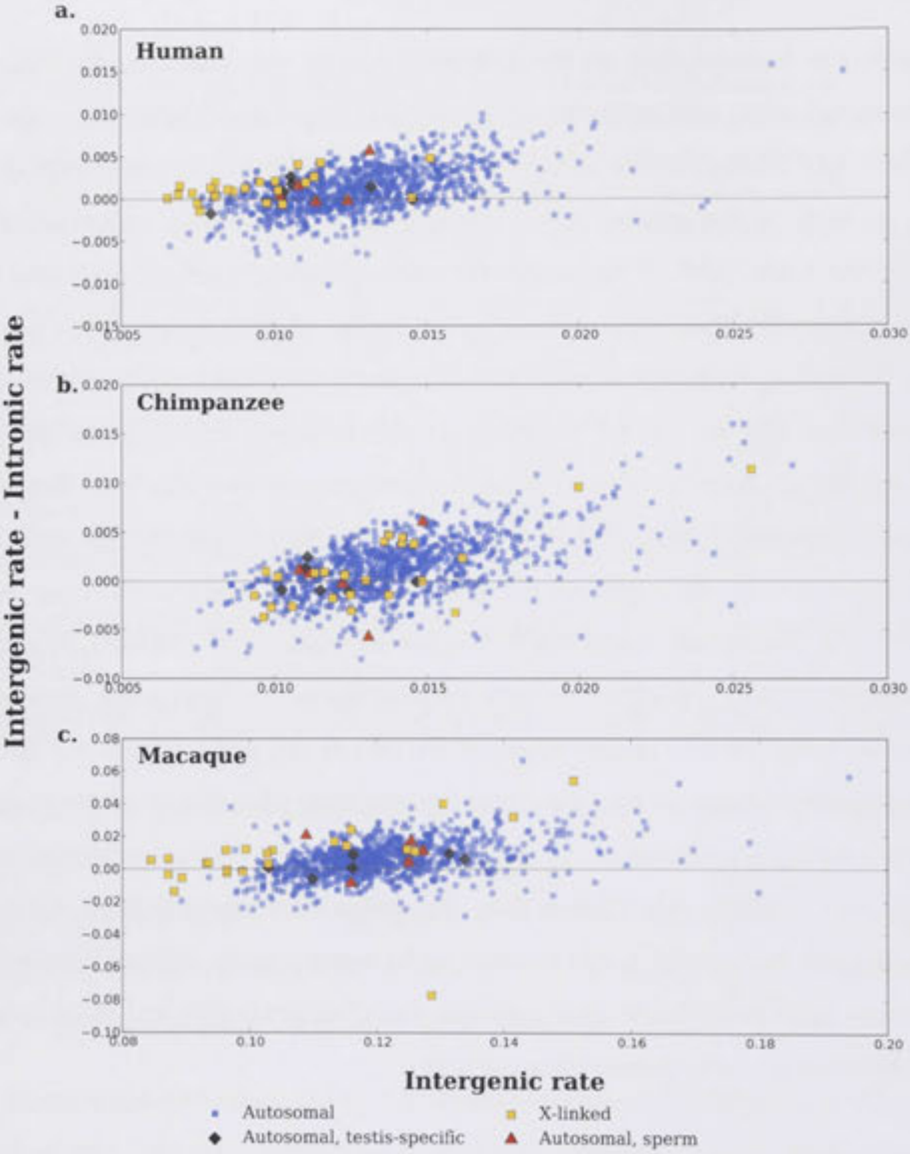

Figure 5.1: The effect of transcription on substitution rate depends on the background substitution rate. Each data point represents data from a flanking intronic and intergenic alignment pair. The branch length for the intergenic alignment is contrasted with the difference between this length and the equivalent length estimate from the intronic alignment. The 'testis-specific' genes are expressed specifically within the testis in humans. The 'sperm' genes are involved in spermatogenesis in humans.

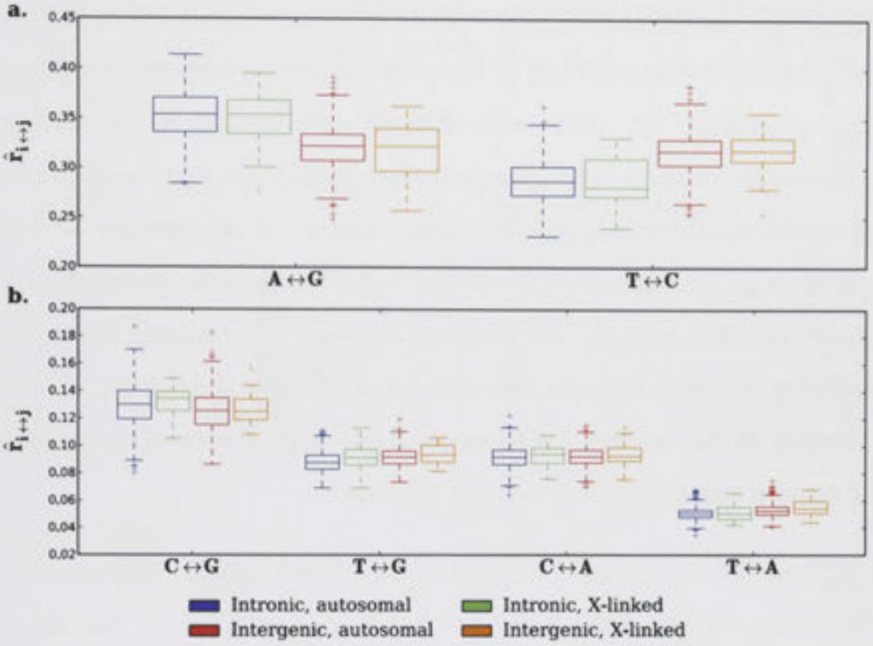

Figure 5.2: Comparison of strand asymmetry in intronic and integenic GTR parameter estimates in the flanking pair data set. a. Transitions and **b.** transversions. Each set of GTR parameters was scaled to sum to one. Whiskers and outliers are defined as in Figure 3.1.

more similar in intergenic than in intronic data (Figure 5.2). No prominent differences in the distributions of GTR parameter estimates for autosomal and X linked alignments were found in either the flanking intronic or intergenic data sets. A slight reduction of the instantaneous, relative $T \leftrightarrow G$ substitution rate in autosomal compared with X-linked data was observed (Figure 5.2 b.), but as this pattern of asymmetry was common to intronic and intergenic data it is not obviously related to transcription.

As expected from the analysis of strand asymmetry for single nucleotide substitutions, strand asymmetry in dinucleotide substitution rate estimates was much stronger (Figure 5.3) and much more consistent (Table 5.2) in intronic than in intergenic data. The "strength" of strand asymmetry amongst the alignments in a data set was defined as the log of the average ratio of strand complementary parameter estimates. The ratio was defined with respect to the dinucleotide that produced an average ratio greater than one, and the log was taken because ratios spanned a narrow range. The "consistency" of asymmetry was defined as the percentage of alignments for which one dinucleotide parameter estimate was greater than the estimate for the complementary parameter. If the direction of asymmetry is not consistent, the consistency score should be 50% and the asymmetry score should be 1, as the member of a complementary parameter pair with the greater rate estimate should be random.

Strand asymmetry for X-linked and autosomal alignments is compared in Figure 5.3. The size of each circle in Figure 5.3 is proportional to the strength of strand asymmetry and circles are scaled with respect to the smallest asymmetry estimate, which occurred in the intergenic data set. Values are shown with respect to the member of a complementary dinucleotide pair with the greater substitution rate estimate for the majority of alignments in each data set. Asymmetry estimates for X-linked and autosomal alignments are overlaid. In some locations on Figure 5.3, only one asymmetry strength estimate is displayed. In these cases, the direction of asymmetry differed between the X-linked and autosomal data sets. For example, consider the lower triangle of Figure 5.3, which shows results from the flanking intronic data set. For autosomal alignments, $\hat{r}_{CG \leftrightarrow CC}$ estimates tended to be greater than the complementary $\hat{r}_{GG \leftrightarrow CG}$ estimates, so the asymmetry strength value for the autosomal alignments is shown at row 'CG' and column 'CC'. The opposite was true of X-linked intronic data, so the asymmetry strength value for X-linked data is shown in row 'GG' and column 'CG'.

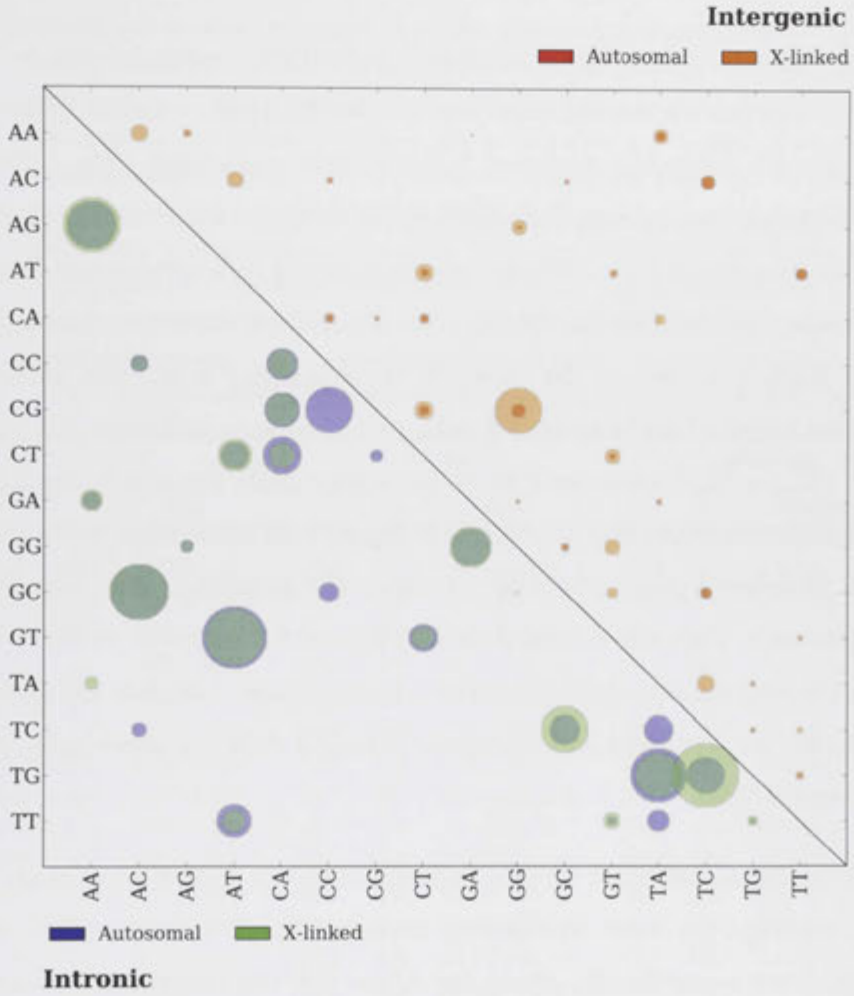

Figure 5.3: Strength of strand asymmetry in complementary dinucleotide substitution rates for flanking intronic and intergenic alignments. The log of the mean ratio of complementary dinucleotide parameters was used as a measure of strand asymmetry, and the size of each bubble is proportional to $log(\frac{D_1}{D_2})$, where $\mathbf{D_1}$ is defined as the dinucleotide substitution that results in an value greater than 1, and the entry for each complimentary parameter pair is displayed with respect to $\mathbf{D_1}$. In the cases where different members of a dinucleotide pair were greater in X-linked and autosomal data, X-linked and autosomal values are still displayed with respect to $\mathbf{D_1}$ as defined above, and displayed in different locations on the plot. The substitution rate difference between complementary dinucleotide parameters is defined as $(r_{GTR_1} * r_{D_1}) - (r_{GTR_2} * r_{D_2})$. The size of each circle is proportional to the strength of asymmetry.

Parameter pairs with strongly asymmetric substitution rates tended to have consistently asymmetric substitution rates, but the direction of strand asymmetry was not always consistent between X-linked and autosomal data. The six most consistently asymmetric substitutions in intronic data were transitions, with purine transitions ($A \leftrightarrow G$) most often occurring at a greater rate than the complementary pyrimidine transitions ($T \leftrightarrow C$) on the transcribed strand (Table 5.2). For these parameters, the strength of asymmetry was fairly consistent between the X-linked and autosomal data, as can be seen in Figure 5.3. $TC \leftrightarrow TG^{*}$ and $TC \leftrightarrow GC^{*}$ were notable as parameter pairs for which asymmetry was both more consistent and stronger in X-linked than autosomal, intronic data. The most prominent parameter pair for which the autosomal data was visibly more asymmetric than the X-linked data in Figure 5.3 was $CG \leftrightarrow CC^{*}$. This pair also showed the strongest asymmetry in intergenic data, but the direction of asymmetry was reversed and intergenic X-linked data was more asymmetric than autosomal data.

Male-bias in the substitution process was unrelated to strand asymmetry. For example, amongst the most consistently asymmetric substitutions pairs, male-bias varied from exceptionally strong for $AT \leftrightarrow GT^{*}$, to exceptionally weak for $AA \leftrightarrow AG^{*}$ (Table 5.2). This difference cannot be attributed to differences in background substitution patterns. Male-bias for $AT \leftrightarrow GT$ substitutions was increased relative to the intergenic data, but decreased relative to the intergenic data for $AA \leftrightarrow AG$ substitutions.

Thus, although consistent differences in the substitution process of X-linked and autosomal alignments exist, these differences can not be summarised simply in terms of increased or decreased substitution rate asymmetry. In a simple scenario,

Table 5.2: Consistency of strand asymmetry in intronic and flanking intergenic data. Complementary pairs of dinucleotide parameters (**D_1** and **D_2**) are sorted from most to least consistent in intronic, autosomal data, where consistent refers to how often one parameter estimate was greater than the complimentary parameter estimate. **A%** and **X%** are the percentages of autosomal and X-linked alignments respectively where the instantaneous, relative rate of $\mathbf{D_1}$, was greater than $\mathbf{D_2}$. The substitution rate estimates used to rank dinucleotide parameter pairs were comprised of the product of a GTR and a dinucleotide term. Values were estimated for the flanking pair data set using the set of complementary dinucleotide model set. The **Sex-bias (Ranks)** columns show estimates of male-bias for each parameter pair, and the ranks of these estimates with respect to all dinucleotide substitutions. These estimates were calculated using the flanking pair data set, and so differ from the estimates of male bias for intronic and intergenic data presented in Chapter 4.

		Intronic			Intergenic		
D_1	**D_2**	**A %**	**X %**	**α (Ranks)**	**A%**	**X%**	**α (Ranks)**
AT ↔ GT	**AT ↔ AC**	92.6	83.3	1.8, 1.6 (1, 2)	55.8	41.7	1.6, 1.3 (1, 5)
TA ↔ TG	**TA ↔ CA**	91.7	83.3	1.4, 1.1 (5, 12)	54.3	45.8	1.3, 1.2 (4, 8)
AC ↔ GC	**GT ↔ GC**	89.7	87.5	1.0, 1.0 (23, 14)	52.8	33.3	1.3, 0.8 (3, 31)
AA ↔ AG	**TT ↔ CT**	87.8	91.7	0.6, 0.7 (45, 42)	53.7	50.0	0.8, 0.7 (33, 41)
CA ↔ CG	**TG ↔ CG**	83.6	87.5	0.9, 0.8 (29, 33)	57.2	45.8	1.1, 1.0 (14, 21)
GA ↔ GG	**TC ↔ CC**	79.9	79.2	1.0, 1.0 (22, 17)	53.0	41.7	0.9, 0.7 (29 , 38)
TC ↔ TG	**CA ↔ GA**	72.5	79.2	0.7, 1.3 (40, 6)	48.9	41.7	1.0, 0.9 (22, 25)
AA ↔ GA	**TT ↔ TC**	68.2	70.8	1.0, 1.0 (20, 13)	52.7	41.7	1.0, 1.0 (17, 23)
CT ↔ CA	**TG ↔ AG**	66.7	66.7	1.6, 1.2 (3, 9)	50.1	45.8	0.7, 0.9 (40, 27)
TT ↔ AT	**AT ↔ AA**	66.7	62.5	0.8, 0.7 (31, 41)	54.2	50.0	0.8, 0.5 (35, 47)
CT ↔ GT	**AC ↔ AG**	66.3	50.0	1.0, 0.7 (21, 37)	49.2	62.5	0.7, 1.0 (42, 19)
CC ↔ CA	**TG ↔ GG**	66.1	70.8	0.9, 0.8 (27, 34)	51.5	41.7	0.9, 1.1 (28, 18)
CT ↔ AT	**AT ↔ AG**	65.5	70.8	0.9, 1.0 (28, 16)	52.2	62.5	1.2, 0.8 (9, 30)
CC ↔ CG	**CG ↔ GG**	64.5	50.0	0.8, 0.5 (30, 46)	53.6	33.3	0.4, 0.9 (48, 26)
TC ↔ TA	**TA ↔ GA**	64.6	54.2	0.9, 0.7 (26, 43)	48.9	54.2	1.1, 0.7 (12, 37)
AG ↔ GG	**CT ↔ CC**	61.7	58.3	1.2, 1.2 (11, 10)	54.0	66.7	1.5, 1.1 (2, 10)
TC ↔ GC	**GC ↔ GA**	61.2	75.0	0.8, 1.3 (35, 7)	51.2	50.0	0.6, 0.8 (46, 36)
TT ↔ TA	**TA ↔ AA**	58.9	45.8	1.3, 0.8 (8, 32)	47.9	54.2	1.3, 1.0 (7, 20)
CC ↔ AC	**GT ↔ GG**	56.0	45.8	0.8, 0.7 (36, 44)	48.4	37.5	1.3, 1.1 (6, 13)
CC ↔ GC	**GC ↔ GG**	54.4	66.7	0.7, 0.7 (39, 38)	49.8	50.0	1.1, 1.1 (11, 16)
CG ↔ AG	**CT ↔ CG**	52.3	54.2	0.3, 0.3 (47, 48)	51.4	54.2	0.6, 1.1 (45, 15)
TT ↔ GT	**AC ↔ AA**	51.6	45.8	0.9, 1.0 (24, 19)	49.4	29.2	0.8, 0.8 (32, 34)
CA ↔ AA	**TT ↔ TG**	51.5	37.5	1.0, 0.9 (15, 25)	49.5	54.2	0.7, 1.0 (39, 24)
TC ↔ AC	**GT ↔ GA**	50.4	37.5	1.5, 1.0 (4, 18)	51.2	54.2	0.7, 0.7 (43, 44)

X-linked and autosomal data would evolve with the same pattern of asymmetry at different rates, i.e. either the X-linked alignments would have greater strand asymmetry than the autosomal alignments across all complementary parameter pairs, or vice versa. The results in Figure 5.3 are suggestive not just of differences in the extent to which transcription affects the substitution rate of X-linked and autosomal alignments, but also differences in how transcription affects the substitution process.

5.4 Discussion

5.4.1 The effect of transcription on the substitution process is modulated by regional factors.

Comparison of the substitution patterns of neighbouring intronic and intergenic alignments indicated that transcription can have either mutagenic or protective effects, and the effect of transcription on a specific region is related to the substitution process of the surrounding region. When the substitution rate of the surrounding region was relatively slow, reflecting accurate repair or infrequent damage, transcription increased the substitution rate. As the substitution rate of the surrounding region increased, transcription tended to decrease the substitution rate. For transcription to be mutagenic despite the activity of TCR, it must either create lesions that are not repaired or prevent the repair of lesions that would be effectively repaired in non-transcribed DNA. An association between transcription intensity and mutation, but not lesion induction (Kim et al., 2007, Hendriks et al., 2008), suggests that the transcription apparatus blocks the repair of naturally-occurring lesions. In regions where lesions are not effectively repaired, this effect of transcription would not be apparent. The results of this study are consistent with a scenario where transcription inhibits repair in regions which are otherwise effectively repaired, but increases repair in regions which are otherwise poorly repaired. Modulation of the influence of transcription on the substitution rate by regional factors may explain why previous studies have found varying effects of transcription on mutation and substitution.

In the species examined, $\hat{\alpha}$ estimates for intronic and intergenic alignments were most similar when the alignments were sampled from neighbouring regions. That the intergenic alignments located close to an intronic alignment showed the strongest male-bias may be a consequence of the effects of transcription on the substitution rate extending beyond the boundaries of genes. This is further implied by the smaller $\hat{\alpha}$ estimates for intronic sequences located close to a long intergenic region than for intronic sequences located in presumably gene-dense regions where no intergenic sequence alignment was found.

The effect of transcription on the substitution rate, and the relationship between the background (intergenic) and intronic substitution rates differed between X-linked and autosomal data. Transcription was more likely to have a protective effect in X-linked than autosomal human and macaque sequences, when their neighbouring intergenic regions evolved at similar rates. This is consistent with the hypothesis of greater repair in the female than the male germline. Another possibility is that the decline of TCR during spermatogenesis (Xu et al., 2005) causes transcription in males to have mutagenic consequences, as transcription can hinder repair when not coupled with recombination (Li and Smerdon, 2004). If transcription causes more mutation in the male than the female germline, testis-specific genes might be expected have the greatest substitution rates compared to their flanking sequences as they would only be expressed in the mutagenic male germ-cell environment. No such difference in the effect of transcription on male-specific and other genes was apparent.

5.4.2 Male bias in transcribed regions is associated with multiple differences in the substitution process

It was hypothesised that the increased male-bias observed in putatively transcribed sequence alignments was a resultf of TCR, and consequently that differences in male-bias between intronic and neighbouring intergenic alignments would be related to differences in the extent of strand asymmetry in their substitution processes. If male-bias is the result of greater TCR in the female than the male germline, then X-linked sequences, which are most frequently located and presumably most frequently expressed in the female germline, should evolve according to a more asymmetric substitution process than autosomal alignments. However, the results of this study indicate that the effect of transcription on the substitution process cannot be adequately summarised in terms of 'more' or 'less' strand asymmetry. The strength of transcription-associated substitution rate asymmetry is known to differ between different substitution types (Green et al., 2003). The results presented here demonstrate that, at least for comparisons of X-linked and autosomal alignments, the strength of substitution rate asymmetry also does not covary between substitution types: some types of substitutions were equally asymmetric in X-linked and autosomal alignments, but others were more, or less, asymmetric.

As transcription had both positive and negative effects on the substitution rate, it is unsurprising that the effect of transcription on the substitution process was likewise heterogenous. The effects of transcription on nucleotide substitution may be caused by several processes, some of which act differently in male and female germlines. Some of the effects of transcription may also be caused by mechanisms that do not result in strand asymmetry, for example transcription-associated chromatin decondensation. Determining whether the changes in substitution rate asymmetry observed between intronic autosomal and X-linked alignments are sufficient to account for the increased male-bias would be a logical next step in understanding how transcription affects the male-bias.

5.4.3 Transcribed regions have greater male-bias

Male-bias was much stronger in intronic than intergenic alignments (Table 5.1), suggesting a major role for transcription in generating male-bias. As the amount of ancestral polymorphism present in a population at speciation is dependent on, amongst other factors, the mutation rate (e.g. Burgess and Yang, 2008), the possibility that the difference in male-bias is an artefact of mutation rate differences between intronic and intergenic data should be considered. Because the X chromosome and the autosomes have different effective population sizes, they are differentially affected by ancestral polymorphism. Differences in divergence times may therefore over- or underestimate the true male-bias in the mutation rate, depending on the chromosomes compared (see Section 1.0.2 in Chapter 1). Theoretically, substitution rate differences between intronic and intergenic data could affect the extent to which ancestral polymorphism contributes to their divergence and therefore affect male-bias. However, Ebersberger et al.

(2007) found the percentage of sites in aligned primate sequences that show phylogenetic inconsistency, a possible consequence of ancestral polymorphism, did not differ between genic and intergenic regions.

The contribution of transcription to male-bias may explain at least some of the variation in previous estimates of male-bias. In particular, Bohossian et al. (2000) and Lander (2001) both found a very low male-bias for humans using intergenic data, whilst other studies have found a greater male-bias by comparing duplicated genes (see Table 1.2 in Chapter 1). Although estimates of human-specific male-bias calculated in this study for intergenic data was greater than in these previous studies, estimates for intronic data were also higher than in many previous studies. Other factors, such as alignment method (Smith and Hurst, 1999), can also contribute to systematic variation in male-bias estimates between studies.

5.4.4 Male-bias in chimpanzee sequences can be entirely explained by transcription

Branch-specific estimates of male-bias did not show a generation time effect. In both intronic and intergenic data, a strong male-bias was estimated for human branches, with a weaker bias for macaque and very weak bias for chimpanzee. No male-bias was detected for the chimpanzee branch in intergenic data, suggesting that what bias was observed was entirely due to transcription. The replication origin hypothesis for male-biased mutation predicts that male-bias for different species should vary according to their generation times. Depending upon the average generation time assumed, α statistics for humans, chimpanzees and macaques are expected to be around 9.7, 6.2 and 2.5 respectively (see Section 13 in Chapter 1 for the derivation of these values).

Male-bias estimates for human and macaque derived from intronic data were on the upper limit of expected values. Equivalent estimates derived from intergenic data were towards the lower limit of expected values. However, male-bias estimates for the chimpanzee were much smaller than expected values, and indeed significantly smaller than the equivalent estimates for macaque, in both intergenic and intronic data. A generation time effect predicts that male-bias estimates in chimpanzee sequences should be greater than estimates for macaque sequences.

The absence of a generation time effect for estimates of male-bias is not likely to be a consequence of ancestral polymorphism. As ancestral polymorphism is predicted to contribute less to the divergence of X-linked (d_X) than autosomal (d_A) alignments, it should result in estimates of $\frac{d_X}{d_A}$ overestimating the ratio of mutation rates (reviewed in Box 1 of Presgraves and Yi, 2009). The proportional contribution of ancestral polymorphism to sequence divergence decreases with increasing time since speciation, therefore ancestral polymorphism should contribute proportionately less to the divergence of the macaque than the chimpanzee. Estimates of α for macaque would remain greater than estimates for chimpanzee if a correction for ancestral polymorphism was applied.

Previous studies have rarely estimated male-bias for human and chimpanzees separately. A trio of sequences is required to assign differences between the human and chimpanzee sequences to either the human or the chimpanzee branch. Previous studies have often used distance measures that only enable a male-bias estimate to be derived for the combined human-chimpanzee branch.

As the male-bias estimated on the human branch in this study was very strong, an estimate of male-bias derived by combining the very biased human branch with the weakly biased chimpanzee branch is likely to produce an estimate within the wide range of values predicted by the generation time hypothesis. Branch-specific estimates of male-bias have been calculated by Presgraves and Yi (2009), but the estimates were conditional on the alignments chosen. Using alignments of human, chimpanzee, gorilla and macaque sequences, Presgraves and Yi (2009) found that male-bias is stronger in chimpanzees than in humans. However, male bias was stronger for humans than chimpanzees when estimated using alignments that also included orangutan sequences.

Although the weak male-bias for chimpanzee sequences estimated in this study does not agree with the predictions of the replication origin hypothesis for male-bias, this result is less surprising in light of recent debate concerning the unusual substitution process of human and chimpanzee X-chromosomes (e.g. Patterson et al., 2006, Burgess and Yang, 2008, Wakeley, 2008, Presgraves and Yi, 2009). Primarily, this debate has centered on whether the unusually low divergence of the human and chimpanzee X chromosomes is a result of a shorter time since speciation than the autosomal average, a smaller female ancestral population size, a strong male-mutation bias, or selective sweeps on the X chromosome. The methods developed to simultaneously estimate ancestral population size, mutation rate and speciation time to date have given limited consideration to molecular clock violation (Hobolth et al., 2007, Burgess and Yang, 2008). Error parameters were incorporated in Burgess and Yang's (2008) model to account for violation of the molecular clock and sequencing errors, either of which could lead to branch-specific substitution rates. The difference between the human and chimpanzee branch lengths was thought to reflect sequencing errors rather than violation of the molecular clock, because the inclusion of the error component altered the estimated transition rate. When estimating male-bias, error parameter estimates from an autosomal data set were applied to the X-linked data set. Branch lengths estimated in this study suggest that human and chimpanzee X-linked and autosomal sequences differ in the extent to which the molecular clock is violated. Resolving the 'true' male-bias for human and chimpanzee sequences, and quantifying the contribution of transcription towards this will probably require both a more detailed method for analysing the primate speciation process, using a model that allows lineage-specific substitution processes.

Chapter 6

Conclusions

In this study, the replication origin hypothesis for male-biased mutation was evaluated by estimating the contributions of methylation and transcription to male-biased mutation. Both transcription and methylation were found to contribute substantially to male bias estimates, methylation for reasons possibly unrelated to sex. In developing the model of context-dependent nucleotide substitution used to evaluate the influences of these processes, a methodological bias that affects popular variants of context-dependent substitution models was identified. These findings emphasise the importance of testing the assumptions underlying models and statistics used in phylogenetic inference.

6.0.5 Is α appropriate for estimating male-bias?

A fundamental assumption of Miyata et al.'s (1987) method for estimating male bias is that substitution rate differences between the sex chromosomes and the autosomes are caused by sex differences in mutation rather than chromosome-specific effects. Studies that have tested this assumption have consistently found evidence indicating that chromosome-specific effects contribute to α estimates (e.g. McVean and Hurst, 1997, Smith and Hurst, 1999, Pink et al., 2009). In Chapter 4, it was shown that α estimates are confounded by sequence features. Over 15% of the average substitution rate difference between X-linked and autosomal alignments can be attributed to CpG motif frequency differences.

Another assumption of the α statistic is that there is a single, genome-wide value for the male mutation bias. The results of Chapter 5 illustrate that this assumption is violated. Male bias estimates for intronic alignments were consistently greater than estimates for nearby intergenic alignments. Further, in Chapter 3, negative values for α were calculated even when attempts were made to eliminate confounding sources of substitution rate variation.

The α statistic cannot be considered a test of the replication origin hypothesis, as the assumptions of this hypothesis are built in to its formulation and only very extreme values of α can be considered as evidence against this hypothesis. The genome-wide α statistics estimated for intergenic human and macaque sequences in Chapter 5 are in general agreement with the values predicted by the replication origin hypothesis, however, consideration of regional substitution process variation contradicted rather than supported this hypothesis.

Many corrections are required to make data meet the assumptions of the α statistic. Factors that must be accounted for include ancestral diversity, motif composition, and regional substitution rate heterogeneity. Rather than performing extensive data culling and curating in order to make use of the α statistic, it is simpler and more intuitive to consider substitution rates directly, and estimate the effects of nucleotide composition separately. The ability to separate the effects of motif frequencies from those of substitution rates is the most valuable feature of the new substitution model form (Yap et al., 2010) developed as a result of the bias detailed in Chapter 2 and (Lindsay et al., 2008). This property of the model will be useful in future analyses of the reasons for substitution rate differences between the sex-chromosomes and the autosomes.

6.0.6 Do replication errors cause the majority of nucleotide substitutions?

Several arguments against, and no arguments for a major contribution of replication errors to the substitution process were presented in this study. Most critically, species-specific estimates of α calculated in Chapter 5 clearly violated the generation time effect predicted by the replication origin hypothesis. If it is assumed that replication errors accumulate at a rate proportional to the number of DNA replications, the almost complete absence of a male bias in the chimpanzee lineage would imply that replication is not the major contributor to the mutation rate. The inadequacy of this model for replication errors is further supported by experimental studies (see Section 1.0.4 in Chapter 1), and by several other studies that have also found that variation in α estimates cannot be completely explained by the replication origin hypothesis (e.g. Smith and Hurst, 1999, Pink et al., 2009). Instead, the results of Chapters 4 and 5 support genome-wide and transcription-coupled repair in the female germline as a major contributors to male-biased evolution. Evaluating the contribution of replication to male-biased mutation will require a more accurate understanding of how replication errors accumulate.

Further consideration of the factors that potentially contribute to male bias would be aided by the use of a wider variety of substitution models. In particular, relaxing the modeling assumption of reversibility would be beneficial for evaluating the contributions of recombination and oxidation. Biased gene conversion is thought to increase the rate of substitution to G and C nucleotides, whilst oxidation is believed to primarily cause mutations of guanine nucleotides (Wang et al., 1998). Both of these effects may have been obscured by the non-reversible used in this study for evaluating methylation and transcription. Like methylation and transcription, recombination has predictable effects on the substitution process, and the effects of oxidative damage can potentially be identified by exploiting the sequence context of substitutions (Stoltzfus, 2008). The use of non-reversible substitution models may also help resolve the mechanism of $AT \leftrightarrow GT$ substitutions. The results of this study indicate that $AT \leftrightarrow GT$ substitutions make an important contribution to male bias and warrant further investigation, being important in improving model fit, exhibiting the strongest male bias of all dinucleotide substitutions in intergenic data, and the most substitution rate asymmetry in intronic data.

6.0.7 Implications for understanding the mechanism of mutations

Variation in mutation rates of different nucleotides has previously related to the fidelity and context-preferences of DNA polymerases. An illustrative example is the well-known bias for transitions over transversions. Watson and Crick (1953) proposed that the probability of a nucleotide being misincorporated depended on its ability to form a base pair with the complementary nucleotide in a similar way to the correct nucleotide, preserving the normal helical structure. Topal and Fresco (1976) proposed a scheme whereby the greater rate of transitions compared with transversions is explained by the ability of uncommon tautomeric forms of nucleotides to pair with a standard nucleotide in the same geometric configuration as standard Watson-Crick nucleotide pairs. However, structural analyses have indicated that mismatches typically involve the major tautomeric nucleotide forms (Morgan, 1993). If replication errors do not make a major contribution to the mutation rate, more work is needed before the base substitution spectrum can be understood at the molecular level for the majority of substitutions.

The Q matrix is often given only limited consideration by phylogenetic studies (Oscamou et al., 2008), despite the potential for the processes that most influence evolution to be inferred from Q as attempted in this study. Research into the biological interpretation of a Q matrix, and the effects of model assumptions on Q matrix parameter estimates, will help future studies make better use of the information provided by Q.

Chapter 7

Supplementary material

Table 7.1: Estimates of male-bias for intergenic data, based on comparison of mean values for chromosomes 3,5 and 6 with X Estimates were calculated as described in Table 4.1 Abbreviations are **Ts** - transition, **Tv** - transversion and **CpG** - substitution involving a CpG dinucleotide

Rank	Context	GTR	$\hat{\alpha}_{GTR\times dinuc}$	$\hat{\alpha}_{GTR}$	$\hat{\alpha}_{dinuc}$	Type
1	$AT \leftrightarrow GT$	$A \leftrightarrow G$	1.46 (1.31–1.62)	1.04 (0.98–1.10)	1.40 (1.29–1.53)	Ts
2	$AT \leftrightarrow AC$	$T \leftrightarrow C$	1.32 (1.19–1.46)	1.00 (0.95–1.06)	1.31 (1.21–1.43)	Ts
3	$CC \leftrightarrow GC$	$C \leftrightarrow G$	1.26 (1.06–1.52)	1.07 (0.99–1.15)	1.18 (0.97–1.43)	Tv
4	$TA \leftrightarrow TG$	$A \leftrightarrow G$	1.25 (1.14–1.39)	1.06 (0.99–1.14)	1.18 (1.10–1.27)	Ts
5	$TA \leftrightarrow CA$	$T \leftrightarrow C$	1.21 (1.10–1.35)	1.00 (0.94–1.07)	1.21 (1.11–1.31)	Ts
6	$CA \leftrightarrow GA$	$C \leftrightarrow G$	1.15 (0.99–1.34)	1.07 (0.98–1.15)	1.07 (0.91–1.25)	Tv
7	$TC \leftrightarrow TG$	$C \leftrightarrow G$	1.11 (0.95–1.30)	1.07 (0.99–1.16)	1.05 (0.90–1.24)	Tv
8	$AG \leftrightarrow GG$	$A \leftrightarrow G$	1.10 (1.00–1.21)	1.07 (0.99–1.15)	1.02 (0.91–1.13)	Ts
9	$CT \leftrightarrow CC$	$T \leftrightarrow C$	1.09 (1.00–1.19)	1.01 (0.95–1.07)	1.08 (0.98–1.19)	Ts
10	$AC \leftrightarrow AG$	$C \leftrightarrow G$	1.07 (0.92–1.24)	1.11 (1.02–1.20)	0.96 (0.83–1.13)	Tv
11	$GC \leftrightarrow GG$	$C \leftrightarrow G$	1.06 (0.90–1.29)	1.07 (0.99–1.16)	0.99 (0.81–1.18)	Tv
12	$CC \leftrightarrow CA$	$C \leftrightarrow A$	1.05 (0.89–1.22)	0.92 (0.85–0.98)	1.14 (0.98–1.35)	Tv
13	$TG \leftrightarrow CG$	$T \leftrightarrow C$	1.03 (0.90–1.19)	1.01 (0.95–1.08)	1.01 (0.88–1.16)	CpG Ts
14	$TG \leftrightarrow AG$	$T \leftrightarrow A$	1.02 (0.85–1.22)	0.69 (0.64–0.76)	1.46 (1.23–1.75)	Tv
15	$AC \leftrightarrow GC$	$A \leftrightarrow G$	1.01 (0.89–1.15)	1.07 (1.00–1.14)	0.95 (0.86–1.05)	Ts
16	$CA \leftrightarrow CG$	$A \leftrightarrow G$	1.01 (0.87–1.18)	1.07 (1.00–1.14)	0.95 (0.82–1.09)	CpG Ts
17	$CC \leftrightarrow AC$	$C \leftrightarrow A$	1.01 (0.86–1.20)	0.92 (0.86–0.98)	1.11 (0.96–1.31)	Tv
18	$CT \leftrightarrow GT$	$C \leftrightarrow G$	1.00 (0.86–1.16)	1.11 (1.02–1.21)	0.89 (0.76–1.03)	Tv
19	$TG \leftrightarrow GG$	$T \leftrightarrow G$	0.99 (0.85–1.16)	0.86 (0.80–0.92)	1.15 (0.99–1.33)	Tv
20	$AA \leftrightarrow GA$	$A \leftrightarrow G$	0.97 (0.87–1.07)	1.07 (1.00–1.15)	0.90 (0.81–0.99)	Ts
21	$TT \leftrightarrow TC$	$T \leftrightarrow C$	0.96 (0.88–1.06)	1.01 (0.95–1.09)	0.94 (0.85–1.03)	Ts
22	$GA \leftrightarrow GG$	$A \leftrightarrow G$	0.96 (0.87–1.08)	1.07 (1.00–1.14)	0.90 (0.81–1.00)	Ts
23	$TC \leftrightarrow TA$	$C \leftrightarrow A$	0.96 (0.82–1.12)	0.93 (0.86–0.99)	1.04 (0.89–1.22)	Tv
24	$GC \leftrightarrow GA$	$C \leftrightarrow A$	0.94 (0.79–1.12)	0.93 (0.87–1.00)	1.00 (0.83–1.19)	Tv
25	$GT \leftrightarrow GC$	$T \leftrightarrow C$	0.90 (0.81–1.01)	1.02 (0.96–1.08)	0.89 (0.81–0.97)	Ts
26	$TC \leftrightarrow CC$	$T \leftrightarrow C$	0.89 (0.79–0.98)	1.02 (0.95–1.09)	0.87 (0.79–0.96)	Ts
27	$CT \leftrightarrow AT$	$C \leftrightarrow A$	0.89 (0.77–1.04)	0.93 (0.87–1.00)	0.95 (0.81–1.09)	Tv
28	$GT \leftrightarrow GG$	$T \leftrightarrow G$	0.88 (0.73–1.04)	0.88 (0.82–0.95)	0.99 (0.83–1.17)	Tv
29	$CG \leftrightarrow AG$	$C \leftrightarrow A$	0.87 (0.69–1.09)	0.92 (0.86–0.98)	0.95 (0.75–1.17)	CpG Tv
30	$CG \leftrightarrow GG$	$C \leftrightarrow G$	0.87 (0.66–1.14)	1.10 (1.02–1.20)	0.75 (0.54–0.99)	CpG Tv
31	$AT \leftrightarrow AG$	$T \leftrightarrow G$	0.86 (0.74–0.99)	0.88 (0.82–0.95)	0.97 (0.85–1.13)	Tv
32	$TT \leftrightarrow GT$	$T \leftrightarrow G$	0.85 (0.73–0.97)	0.94 (0.88–1.01)	0.90 (0.78–1.03)	Tv
33	$AC \leftrightarrow AA$	$C \leftrightarrow A$	0.85 (0.73–0.97)	0.94 (0.88–1.01)	0.90 (0.78–1.03)	Tv
34	$TC \leftrightarrow GC$	$T \leftrightarrow G$	0.85 (0.71–1.01)	0.88 (0.82–0.95)	0.95 (0.80–1.16)	Tv
35	$TT \leftrightarrow CT$	$T \leftrightarrow C$	0.84 (0.77–0.93)	1.02 (0.96–1.08)	0.82 (0.75–0.89)	Ts
36	$CT \leftrightarrow CA$	$T \leftrightarrow A$	0.83 (0.70–1.00)	0.69 (0.64–0.76)	1.18 (0.99–1.41)	Tv
37	$AA \leftrightarrow AG$	$A \leftrightarrow G$	0.82 (0.75–0.90)	1.09 (1.02–1.16)	0.75 (0.70–0.82)	Ts
38	$CA \leftrightarrow AA$	$C \leftrightarrow A$	0.82 (0.72–0.95)	0.95 (0.88–1.02)	0.86 (0.75–0.99)	Tv
39	$TT \leftrightarrow TG$	$T \leftrightarrow G$	0.82 (0.72–0.93)	0.89 (0.83–0.95)	0.93 (0.82–1.06)	Tv
40	$TA \leftrightarrow GA$	$T \leftrightarrow G$	0.80 (0.69–0.94)	0.89 (0.83–0.96)	0.91 (0.78–1.06)	Tv
41	$TA \leftrightarrow AA$	$T \leftrightarrow A$	0.80 (0.69–0.93)	0.74 (0.66–0.82)	1.07 (0.92–1.26)	Tv
42	$CC \leftrightarrow CG$	$C \leftrightarrow G$	0.80 (0.60–1.09)	1.10 (1.01–1.19)	0.69 (0.51–0.92)	CpG Tv
43	$CT \leftrightarrow CG$	$T \leftrightarrow G$	0.79 (0.61–1.01)	0.87 (0.80–0.93)	0.90 (0.69–1.14)	CpG Tv
44	$TT \leftrightarrow TA$	$T \leftrightarrow A$	0.72 (0.61–0.86)	0.74 (0.67–0.81)	0.94 (0.79–1.11)	Tv
45	$TC \leftrightarrow AC$	$T \leftrightarrow A$	0.70 (0.55–0.88)	0.74 (0.68–0.81)	0.94 (0.77–1.14)	Tv
46	$GT \leftrightarrow GA$	$T \leftrightarrow A$	0.69 (0.55–0.88)	0.74 (0.68–0.81)	0.94 (0.76–1.16)	Tv
47	$TT \leftrightarrow AT$	$T \leftrightarrow A$	0.69 (0.58–0.82)	0.78 (0.71–0.86)	0.91 (0.78–1.06)	Tv
48	$AT \leftrightarrow AA$	$T \leftrightarrow A$	0.59 (0.49–0.71)	0.78 (0.71–0.85)	0.78 (0.65–0.91)	Tv

Table 7.2: Estimates of $\hat{\alpha}$ for intronic data Estimates were calculated as described in Table 4.1 Abbreviations are **Ts** - transition, **Tv** - transversion and **CpG** - substitution involving a CpG dinucleotide

Rank	Context	GTR	$\hat{\alpha}_{GTR \times dinuc}$	$\hat{\alpha}_{GTR}$	$\hat{\alpha}_{dinuc}$	Type
1	$AT \leftrightarrow GT$	$A \leftrightarrow G$	1.59 (1.35–1.90)	1.05 (0.96–1.13)	1.53 (1.36–1.74)	Ts
2	$TA \leftrightarrow TG$	$A \leftrightarrow G$	1.39 (1.19–1.64)	1.06 (0.97–1.16)	1.32 (1.16–1.50)	Ts
3	$TG \leftrightarrow AG$	$T \leftrightarrow A$	1.35 (1.01–1.81)	0.91 (0.80–1.04)	1.53 (1.16–2.01)	Tv
4	$AT \leftrightarrow AC$	$T \leftrightarrow C$	1.27 (1.09–1.50)	1.04 (0.95–1.12)	1.23 (1.09–1.40)	Ts
5	$CT \leftrightarrow CA$	$T \leftrightarrow A$	1.26 (0.97–1.66)	0.91 (0.79–1.04)	1.43 (1.10–1.85)	Tv
6	$TC \leftrightarrow AC$	$T \leftrightarrow A$	1.23 (0.89–1.70)	0.95 (0.83–1.10)	1.26 (0.87–1.76)	Tv
7	$TT \leftrightarrow TC$	$T \leftrightarrow C$	1.19 (1.02–1.39)	0.99 (0.91–1.10)	1.18 (1.00–1.41)	Ts
8	$CT \leftrightarrow CC$	$T \leftrightarrow C$	1.19 (1.02–1.38)	1.00 (0.91–1.10)	1.21 (1.04–1.39)	Ts
9	$CC \leftrightarrow CA$	$C \leftrightarrow A$	1.15 (0.88–1.48)	0.97 (0.86–1.10)	1.18 (0.92–1.51)	Tv
10	$TA \leftrightarrow CA$	$T \leftrightarrow C$	1.14 (0.99–1.31)	1.02 (0.94–1.11)	1.12 (0.99–1.26)	Ts
11	$TT \leftrightarrow AT$	$T \leftrightarrow A$	1.13 (0.89–1.45)	0.99 (0.86–1.13)	1.14 (0.88–1.48)	Tv
12	$AG \leftrightarrow GG$	$A \leftrightarrow G$	1.12 (0.97–1.30)	1.09 (1.00–1.21)	1.02 (0.87–1.20)	Ts
13	$TC \leftrightarrow TA$	$C \leftrightarrow A$	1.11 (0.86–1.38)	0.98 (0.87–1.11)	1.14 (0.89–1.44)	Tv
14	$AA \leftrightarrow GA$	$A \leftrightarrow G$	1.06 (0.92–1.24)	1.09 (1.00–1.20)	0.97 (0.84–1.12)	Ts
15	$TG \leftrightarrow GG$	$T \leftrightarrow G$	1.05 (0.78–1.44)	0.86 (0.76–0.98)	1.25 (0.96–1.64)	Tv
16	$TC \leftrightarrow TG$	$C \leftrightarrow G$	1.04 (0.80–1.32)	0.82 (0.71–0.93)	1.23 (0.94–1.60)	Tv
17	$CA \leftrightarrow AA$	$C \leftrightarrow A$	1.03 (0.81–1.31)	0.98 (0.87–1.11)	1.06 (0.83–1.33)	Tv
18	$TC \leftrightarrow GC$	$T \leftrightarrow G$	0.99 (0.77–1.25)	0.87 (0.76–1.00)	1.15 (0.91–1.47)	Tv
19	$GT \leftrightarrow GC$	$T \leftrightarrow C$	0.98 (0.83–1.16)	1.01 (0.93–1.10)	0.96 (0.81–1.11)	Ts
20	$GT \leftrightarrow GA$	$T \leftrightarrow A$	0.97 (0.69–1.35)	0.95 (0.83–1.08)	1.01 (0.73–1.44)	Tv
21	$AC \leftrightarrow GC$	$A \leftrightarrow G$	0.95 (0.79–1.14)	1.09 (1.00–1.20)	0.88 (0.75–1.01)	Ts
22	$CC \leftrightarrow GC$	$C \leftrightarrow G$	0.95 (0.71–1.25)	0.87 (0.76–0.99)	1.07 (0.80–1.44)	Tv
23	$CT \leftrightarrow AT$	$C \leftrightarrow A$	0.92 (0.74–1.15)	1.00 (0.88–1.13)	0.92 (0.73–1.15)	Tv
24	$TA \leftrightarrow AA$	$T \leftrightarrow A$	0.91 (0.69–1.20)	1.06 (0.90–1.22)	0.84 (0.64–1.11)	Tv
25	$CC \leftrightarrow AC$	$C \leftrightarrow A$	0.88 (0.65–1.23)	1.00 (0.88–1.13)	0.87 (0.61–1.23)	Tv
26	$GC \leftrightarrow GA$	$C \leftrightarrow A$	0.87 (0.62–1.19)	1.01 (0.89–1.15)	0.87 (0.63–1.19)	Tv
27	$TC \leftrightarrow CC$	$T \leftrightarrow C$	0.86 (0.72–1.01)	1.02 (0.93–1.12)	0.86 (0.75–0.99)	Ts
28	$AC \leftrightarrow AA$	$C \leftrightarrow A$	0.86 (0.66–1.12)	1.01 (0.89–1.16)	0.83 (0.62–1.08)	Tv
29	$GC \leftrightarrow GG$	$C \leftrightarrow G$	0.86 (0.63–1.21)	0.87 (0.76–0.98)	0.97 (0.69–1.34)	Tv
30	$CT \leftrightarrow GT$	$C \leftrightarrow G$	0.85 (0.68–1.05)	0.91 (0.80–1.02)	0.92 (0.72–1.15)	Tv
31	$GT \leftrightarrow GG$	$T \leftrightarrow G$	0.84 (0.62–1.13)	0.88 (0.77–1.00)	0.95 (0.72–1.22)	Tv
32	$GA \leftrightarrow GG$	$A \leftrightarrow G$	0.83 (0.69–0.97)	1.10 (1.01–1.21)	0.73 (0.59–0.90)	Ts
33	$TA \leftrightarrow GA$	$T \leftrightarrow G$	0.85 (0.65–1.10)	0.88 (0.78–1.01)	0.97 (0.74–1.24)	Tv
34	$TT \leftrightarrow GT$	$T \leftrightarrow G$	0.84 (0.67–1.04)	0.88 (0.77–1.01)	0.96 (0.78–1.19)	Tv
35	$CA \leftrightarrow CG$	$A \leftrightarrow G$	0.81 (0.64–1.00)	1.08 (0.99–1.19)	0.75 (0.60–0.93)	CpG Ts
36	$TT \leftrightarrow TG$	$T \leftrightarrow G$	0.81 (0.65–1.03)	0.89 (0.76–1.02)	0.88 (0.69–1.14)	Tv
37	$CA \leftrightarrow GA$	$C \leftrightarrow G$	0.79 (0.59–1.07)	0.82 (0.72–0.93)	0.97 (0.75–1.29)	Tv
38	$TT \leftrightarrow TA$	$T \leftrightarrow A$	0.78 (0.60–0.99)	1.06 (0.91–1.23)	0.71 (0.53–0.92)	Tv
39	$AT \leftrightarrow AG$	$T \leftrightarrow G$	0.78 (0.58–1.01)	0.88 (0.77–1.00)	0.89 (0.70–1.13)	Tv
40	$AA \leftrightarrow AG$	$A \leftrightarrow G$	0.76 (0.66–0.90)	1.10 (1.00–1.20)	0.68 (0.58–0.80)	Ts
41	$AT \leftrightarrow AA$	$T \leftrightarrow A$	0.76 (0.59–0.97)	0.99 (0.86–1.13)	0.77 (0.59–0.99)	Tv
42	$TT \leftrightarrow CT$	$T \leftrightarrow C$	0.71 (0.61–0.84)	1.04 (0.95–1.14)	0.68 (0.57–0.81)	Ts
43	$TG \leftrightarrow CG$	$T \leftrightarrow C$	0.67 (0.53–0.85)	1.02 (0.93–1.10)	0.67 (0.53–0.84)	CpG Ts
44	$CC \leftrightarrow CG$	$C \leftrightarrow G$	0.67 (0.44–0.97)	0.85 (0.73–0.97)	0.75 (0.52–1.06)	CpG Tv
45	$AC \leftrightarrow AG$	$C \leftrightarrow G$	0.65 (0.48–0.85)	0.91 (0.79–1.02)	0.72 (0.54–0.94)	Tv
46	$CG \leftrightarrow AG$	$C \leftrightarrow A$	0.58 (0.35–0.91)	0.97 (0.85–1.10)	0.59 (0.34–0.90)	CpG Tv
47	$CT \leftrightarrow CG$	$T \leftrightarrow G$	0.55 (0.32–0.83)	0.86 (0.75–0.98)	0.64 (0.42–0.94)	CpG Tv
48	$CG \leftrightarrow GG$	$C \leftrightarrow G$	0.43 (0.20–0.76)	0.85 (0.73–0.97)	0.52 (0.28–0.88)	CpG Tv

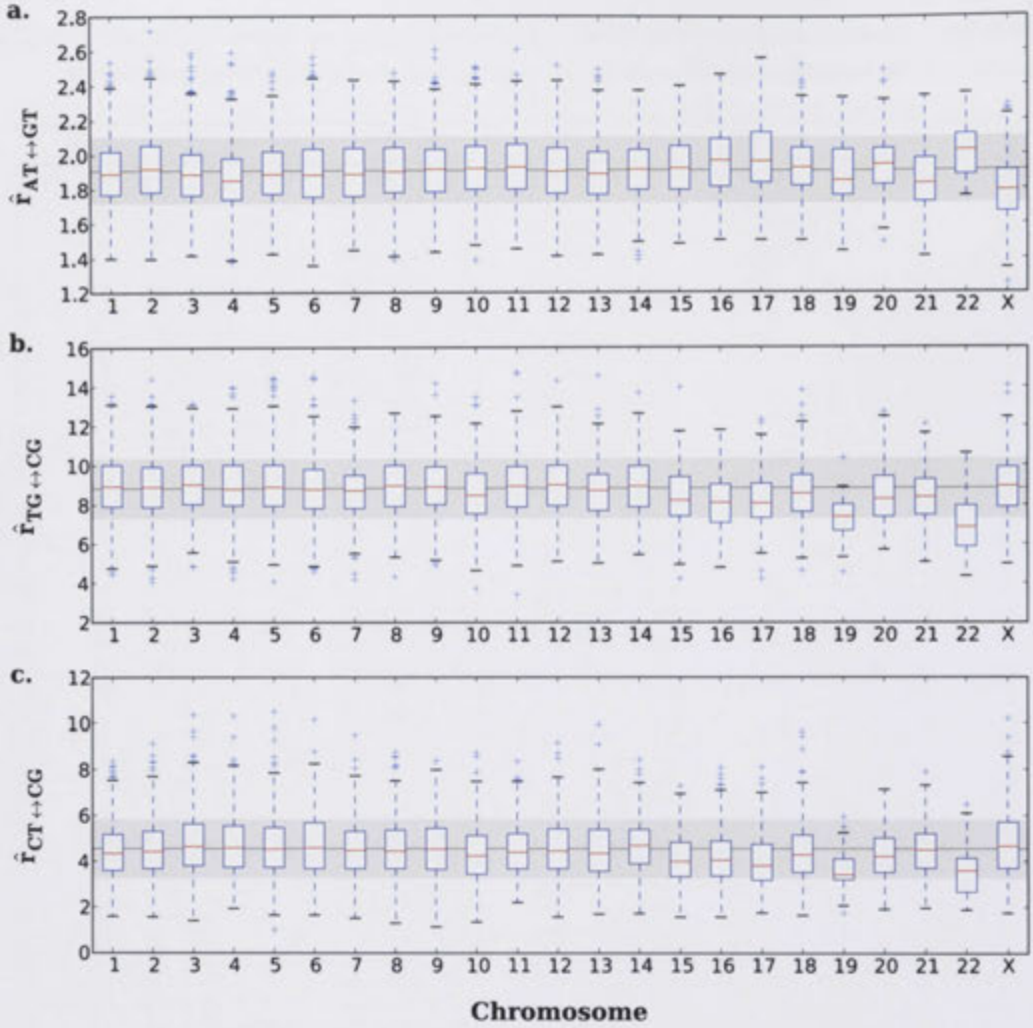

Figure 7.1: Dinucleotide substitution rate estimates by chromosome. Estimates are shown for **a.** $AT \leftrightarrow GT$, the most male-biased dinucleotide substitution **b.** $TG \leftrightarrow CG$, the transition caused by deamination of methylated cytosine and **c.** $CT \leftrightarrow CG$, the most female-biased dinucleotide substitution. Dinucleotide substitution rate estimates indicate the extent to which the substitution within the specified dinucleotide context differs from the relevant GTR estimate. The black line behind the box plots is the mean substitution rate estimate for the autosomes, and the area that falls within one standard deviation of this value is shaded grey. An extreme value that differed from the chromosomal mean by more than 5 standard deviations was excluded from panels **b** and **c**. Whiskers and outliers are defined as in Figure 3.1.

Bibliography

Agulnik, A. I., Bishop, C. E., Lerner, J. L., Agulnik, S. I. and Solovyev, V. V. (1997), 'Analysis of mutation rates in the SMCY/SMCX genes shows that mammalian evolution is male driven.', *Mamm Genome* **8**(2), 134–8.

Aitken, R. and De Iuliis, G. (2009), 'On the possible origins of DNA damage in human spermatozoa.', *Mol Hum Reprod* .

Aitken, R. J., De Iuliis, G. N. and McLachlan, R. I. (2009), 'Biological and clinical significance of DNA damage in the male germ line.', *Int J Androl* **32**(1), 46–56.

Allegrucci, C., Thurston, A., Lucas, E. and Young, L. (2005), 'Epigenetics and the germline.', *Reproduction* **129**(2), 137–149.

Allen, J. W., Ehling, U. H., Moore, M. M. and Lewis, S. E. (1995), 'Germ line specific factors in chemical mutagenesis.', *Mutat Res* **330**(1-2), 219–231.

Antonarakis, S. E., Krawczak, M. and Cooper, D. N. (2000), 'Disease-causing mutations in the human genome.', *Eur J Pediatr* **159 Suppl 3**, S173–8.

Aoki, V. W., Emery, B. R., Liu, L. and Carrell, D. T. (2006), 'Protamine levels vary between individual sperm cells of infertile human males and correlate with viability and DNA integrity.', *J Androl* **27**(6), 890–898.

Arndt, P. (2006), 'Reconstruction of ancestral nucleotide sequences and estimation of substitution frequencies in a star phylogeny.', *Gene* .

Arndt, P. F., Burge, C. B. and Hwa, T. (2003), 'DNA sequence evolution with neighbor-dependent mutation.', *J Comput Biol* **10**(3-4), 313–22.

Arndt, P. F. and Hwa, T. (2005), 'Identification and measurement of neighbor-dependent nucleotide substitution processes.', *Bioinformatics* **21**(10), 2322–8.

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**(6), 748–63.

Arndt, P. F., Petrov, D. A. and Hwa, T. (2003), 'Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation.', *Mol Biol Evol* **20**(11), 1887–96.

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

Axelsson, E., Smith, N. G. C., Sundstrom, H., Berlin, S. and Ellegren, H. (2004), 'Male-biased mutation rate and divergence in autosomal, Z-linked and W-linked introns of chicken and turkey.', *Mol Biol Evol* **21**(8), 1538–47.

Baarends, W. M., van der Laan, R. and Grootegoed, J. A. (2001), 'DNA repair mechanisms and gametogenesis.', *Reproduction* **121**(1), 31–39.

Baele, G., Van de Peer, Y. and Vansteelandt, S. (2008), 'A model-based approach to study nearest-neighbor influences reveals complex substitution patterns in non-coding sequences.', *Syst Biol* **57**(5), 675–692.

Baisnee, P.-F., Hampson, S. and Baldi, P. (2002), 'Why are complementary DNA strands symmetric?', *Bioinformatics* **18**(8), 1021–1033.

Bartosch-Harlid, A., Berlin, S., Smith, N. G. C., Moller, A. P. and Ellegren, H. (2003), 'Life history and the male mutation bias.', *Evolution Int J Org Evolution* **57**(10), 2398–406.

Becker, J., Schwaab, R., Moller-Taube, A., Schwaab, U., Schmidt, W., Brackmann, H. H., Grimm, T., Olek, K. and Oldenburg, J. (1996), 'Characterization of the factor VIII defect in 147 patients with sporadic hemophilia a: family studies indicate a mutation type-dependent sex ratio of mutation frequencies.', *Am J Hum Genet* **58**(4), 657–670.

Berlin, S., Brandstrom, M., Backstrom, N., Axelsson, E., Smith, N. G. C. and Ellegren, H. (2006), 'Substitution rate heterogeneity and the male mutation bias.', *J Mol Evol* **62**(2), 226–33.

Bernardi, G. (1993), 'The vertebrate genome: isochores and evolution.', *Mol Biol Evol* **10**(1), 186–204.

Bielawski, J. P., Dunn, K. A. and Yang, Z. (2000), 'Rates of nucleotide substitution and mammalian nuclear gene evolution. approximate and maximum-likelihood methods lead to different conclusions.', *Genetics* **156**(3), 1299–308.

Bill, C. A., Duran, W. A., Miselis, N. R. and Nickoloff, J. A. (1998), 'Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches.', *Genetics* **149**(4), 1935–43.

Blake, R. D., Hess, S. T. and Nicholson-Tuell, J. (1992), 'The influence of nearest neighbors on the rate and pattern of spontaneous point mutations.', *J Mol Evol* **34**(3), 189–200.

Bohossian, H. B., Skaletsky, H. and Page, D. C. (2000), 'Unexpectedly similar rates of nucleotide substitution found in male and female hominids.', *Nature* **406**(6796), 622–5.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. and Weber, J. L. (1998), 'Comprehensive human genetic maps: individual and sex-specific variation in recombination.', *Am J Hum Genet* **63**(3), 861–869.

Brown, T. C. and Jiricny, J. (1988), 'Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells.', *Cell* **54**(5), 705–11.

Burgess, R. and Yang, Z. (2008), 'Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors.', *Mol Biol Evol* **25**(9), 1979–1994.

Cabelof, D. C., Raffoul, J. J., Yanamadala, S., Ganir, C., Guo, Z. and Heydari, A. R. (2002), 'Attenuation of DNA polymerase beta-dependent base excision repair and increased DMS-induced mutagenicity in aged mice.', *Mutat Res* **500**(1-2), 135–145.

Chang, B. H. and Li, W. H. (1995), 'Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked Ube 1 genes and pseudogenes.', *J Mol Evol* **40**(1), 70–7.

Chang, B. H., Shimmin, L. C., Shyue, S. K., Hewett-Emmett, D. and Li, W. H. (1994), 'Weak male-driven molecular evolution in rodents.', *Proc Natl Acad Sci U S A* **91**(2), 827–31.

Chang, J. S. (2009), 'Parental smoking and childhood leukemia.', *Methods Mol Biol* **472**, 103–137.

Choi, S.-K., Yoon, S.-R., Calabrese, P. and Arnheim, N. (2008), 'A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations.', *Proc Natl Acad Sci U S A* **105**(29), 10143–10148.

Chow, J. C., Yen, Z., Ziesche, S. M. and Brown, C. J. (2005), 'Silencing of the mammalian X chromosome.', *Annu Rev Genomics Hum Genet* **6**, 69–92.

Cocuzza, M., Sikka, S. C., Athayde, K. S. and Agarwal, A. (2007), 'Clinical relevance of oxidative stress and sperm chromatin damage in male infertility: an evidence based analysis.', *Int Braz J Urol* **33**(5), 603–621.

Cooper, D. N. and Youssoufian, H. (1988), 'The CpG dinucleotide and human genetic disease.', *Hum Genet* **78**(2), 151–5.

Cortazar, D., Kunz, C., Saito, Y., Steinacher, R. and Schar, P. (2007), 'The enigmatic thymine DNA glycosylase.', *DNA Repair (Amst)* **6**(4), 489–504.

Coulondre, C., Miller, J. H., Farabaugh, P. J. and Gilbert, W. (1978), 'Molecular basis of base substitution hotspots in *Escherichia coli*.', *Nature* **274**(5673), 775–80.

Crow, J. F. (1999), 'Spontaneous mutation in man.', *Mutat Res* **437**(1), 5–9.

Crow, J. F. (2006), 'Age and sex effects on human mutation rates: an old problem with new complexities.', *J Radiat Res (Tokyo)* **47 Suppl B**, B75–82.

Dadoune, J.-P., Siffroi, J.-P. and Alfonsi, M.-F. (2004), 'Transcription in haploid male germ cells.', *Int Rev Cytol* **237**, 1–56.

Datta, A. and Jinks-Robertson, S. (1995), 'Association of increased spontaneous mutation rates with high levels of transcription in yeast.', *Science* **268**(5217), 1616–1619.

Dreszer, T., Wall, G., Haussler, D. and Pollard, K. (2007), 'Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion.', *Genome Res* .

Driscoll, D. J. and Migeon, B. R. (1990), 'Sex difference in methylation of single-copy genes in human meiotic germ cells: implications for X chromosome inactivation, parental imprinting, and origin of CpG mutations.', *Somat Cell Mol Genet* **16**(3), 267–82.

Duret, L. and Arndt, P. F. (2008), 'The impact of recombination on nucleotide substitutions in the human genome.', *PLoS Genet* **4**(5), e1000071.

Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M. and von Haeseler, A. (2007), 'Mapping human genetic ancestry.', *Mol Biol Evol* **24**(10), 2266–2276.

Ebersberger, I., Metzler, D., Schwarz, C. and Paabo, S. (2002), 'Genomewide comparison of DNA sequences between humans and chimpanzees.', *Am J Hum Genet* **70**(6), 1490–7.

Ehmcke, J., Wistuba, J. and Schlatt, S. (2006), 'Spermatogonial stem cells: questions, models and perspectives.', *Hum Reprod Update* **12**(3), 275–282.

Ehrenhofer-Murray, A. E. (2004), 'Chromatin dynamics at DNA replication, transcription and repair.', *Eur J Biochem* **271**(12), 2335–2349.

Eichenlaub-Ritter, U. and Peschke, M. (2002), 'Expression in in-vivo and in-vitro growing and maturing oocytes: focus on regulation of expression at the translational level.', *Hum Reprod Update* **8**(1), 21–41.

El-Maarri, O., Olek, A., Balaban, B., Montag, M., van der Ven, H., Urman, B., Olek, K., Caglayan, S. H., Walter, J. and Oldenburg, J. (1998), 'Methylation levels at selected CpG sites in the factor VIII and FGFR3 genes, in mature female and male germ cells: implications for male-driven evolution.', *Am J Hum Genet* **63**(4), 1001–8.

Ellegren, H. (2000), 'Heterogeneous mutation processes in human microsatellite DNA sequences.', *Nat Genet* **24**(4), 400–2.

Ellegren, H. and Fridolfsson, A. K. (1997), 'Male-driven evolution of DNA sequences in birds.', *Nat Genet* **17**(2), 182–4.

Ellegren, H. and Fridolfsson, A.-K. (2003), 'Sex-specific mutation rates in salmonid fish.', *J Mol Evol* **56**(4), 458–63.

Ellegren, H., Smith, N. G. C. and Webster, M. T. (2003), 'Mutation rate variation in the mammalian genome.', *Curr Opin Genet Dev* **13**(6), 562–8.

Engelhardt, A., Heistermann, M., Hodges, J. K., Nürnberg, P. and Niemitz, C. (2006), 'Determinants of male reproductive success in wild long-tailed macaques (macaca fascicularis) - male monopolisation, female mate choice or post-copulatory mechanisms', *Behav Ecol Sociobiol* **59**, 740–752.

Episkopou, H., Kyrtopoulos, S. A., Sfikakis, P. P., Fousteri, M., Dimopoulos, M. A., Mullenders, L. H. F. and Souliotis, V. L. (2009), 'Association between transcriptional activity, local chromatin structure, and the efficiencies of both subpathways of nucleotide excision repair of melphalan adducts.', *Cancer Res* **69**(10), 4424–4433.

Erlandsson, R., Wilson, J. F. and Paabo, S. (2000), 'Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans.', *Mol Biol Evol* **17**(5), 804–812.

Evans, D. G. R., Maher, E. R. and Baser, M. E. (2005), 'Age related shift in the mutation spectra of germline and somatic NF2 mutations: hypothetical role of dna repair mechanisms.', *J Med Genet* **42**(8), 630–2.

Felsenstein, J. (1981), 'Evolutionary trees from DNA sequences: a maximum likelihood approach.', *J Mol Evol* **17**(6), 368–76.

Feng, J., Drost, J. B., Scaringe, W. A., Liu, Q. and Sommer, S. S. (2002), 'Mutations in the factor IX gene (F9) during the past 150 years have relative rates similar to ancient mutations.', *Hum Mutat* **19**(1), 49–57.

Feng, Z., Hu, W., Komissarova, E., Pao, A., Hung, M.-C., Adair, G. M. and Tang, M.-s. (2002), 'Transcription-coupled DNA repair is genomic context-dependent.', *J Biol Chem* **277**(15), 12777–12783.

Filipski, J. (1987), 'Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells.', *FEBS Lett* **217**(2), 184–6.

Fraga, C. G., Motchnik, P. A., Shigenaga, M. K., Helbock, H. J., Jacob, R. A. and Ames, B. N. (1991), 'Ascorbic acid protects against endogenous oxidative DNA damage in human sperm.', *Proc Natl Acad Sci U S A* **88**(24), 11003–11006.

Francino, M. P. and Ochman, H. (1999), 'Isochores result from mutation not selection.', *Nature* **400**(6739), 30–1.

Frans, E. M., Sandin, S., Reichenberg, A., Lichtenstein, P., Langstrom, N. and Hultman, C. M. (2008), 'Advancing paternal age and bipolar disorder.', *Arch Gen Psychiatry* **65**(9), 1034–1040.

Frit, P., Kwon, K., Coin, F., Auriol, J., Dubaele, S., Salles, B. and Egly, J. M. (2002), 'Transcriptional activators stimulate DNA repair.', *Mol Cell* **10**(6), 1391–1401.

Fryxell, K. J. and Moon, W.-J. (2005), 'CpG mutation rates in the human genome are highly dependent on local GC content.', *Mol Biol Evol* **22**(3), 650–8.

Fuks, F. (2005), 'DNA methylation and histone modifications: teaming up to silence genes.', *Curr Opin Genet Dev* **15**(5), 490–5.

Gaffney, D. J. and Keightley, P. D. (2008), 'Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals.', *BMC Evol Biol* **8**, 265.

Gage, T. B. (1998), 'The comparative demography of primates: with some comments on the evolution of life histories.', *Annu Rev Anthropol* **27**, 197–221.

Galtier, N. and Gouy, M. (1998), 'Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis.', *Mol Biol Evol* **15**(7), 871–9.

Galtier, N., Piganeau, G., Mouchiroud, D. and Duret, L. (2001), 'GC-content evolution in mammalian genomes: the biased gene conversion hypothesis.', *Genetics* **159**(2), 907–11.

Geraldes, A., Basset, P., Gibson, B., Smith, K. L., Harr, B., Yu, H.-T., Bulatova, N., Ziv, Y. and Nachman, M. W. (2008), 'Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes.', *Mol Ecol* **17**(24), 5349–5363.

Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., Remington, K. A., Strausberg, R. L., Venter, J. C., Wilson, R. K., Batzer, M. A., Bustamante, C. D., Eichler, E. E., Hahn, M. W., Hardison, R. C., Makova, K. D., Miller, W., Milosavljevic, A., Palermo, R. E., Siepel, A., Sikela, J. M., Attaway, T., Bell, S., Bernard, K. E., Buhay, C. J., Chandrabose, M. N., Dao, M., Davis, C., Delehaunty, K. D., Ding, Y., Dinh, H. H., Dugan-Rocha, S., Fulton, L. A., Gabisi, R. A., Garner, T. T., Godfrey, J., Hawes, A. C., Hernandez, J., Hines, S., Holder, M., Hume, J., Jhangiani, S. N., Joshi, V., Khan, Z. M., Kirkness, E. F., Cree, A., Fowler, R. G., Lee, S., Lewis, L. R., Li, Z., Liu, Y.-S., Moore, S. M., Muzny, D., Nazareth, L. V., Ngo, D. N., Okwuonu, G. O., Pai, G., Parker, D., Paul, H. A., Pfannkoch, C., Pohl, C. S., Rogers, Y.-H., Ruiz, S. J., Sabo, A., Santibanez, J., Schneider, B. W., Smith, S. M., Sodergren, E., Svatek, A. F., Utterback, T. R., Vattathil, S., Warren, W., White, C. S., Chinwalla, A. T., Feng, Y., Halpern, A. L., Hillier, L. W., Huang, X., Minx, P., Nelson, J. O., Pepin, K. H., Qin, X., Sutton, G. G., Venter, E., Walenz, B. P., Wallis, J. W., Worley, K. C., Yang, S.-P., Jones, S. M., Marra, M. A., Rocchi, M., Schein, J. E., Baertsch, R., Clarke, L., Csuros, M., Glasscock, J., Harris, R. A., Havlak, P., Jackson, A. R., Jiang, H., Liu, Y., Messina, D. N., Shen, Y., Song, H. X.-Z., Wylie, T., Zhang, L., Birney, E., Han, K., Konkel, M. K., Lee, J., Smit, A. F. A., Ullmer, B., Wang, H., Xing, J., Burhans, R., Cheng, Z., Karro, J. E., Ma, J., Raney, B., She, X., Cox, M. J., Demuth, J. P., Dumas, L. J., Han, S.-G., Hopkins, J., Karimpour-Fard, A., Kim, Y. H., Pollack, J. R., Vinar, T., Addo-Quaye, C., Degenhardt, J., Denby, A., Hubisz,

M. J., Indap, A., Kosiol, C., Lahn, B. T., Lawson, H. A., Marklein, A., Nielsen, R., Vallender, E. J., Clark, A. G., Ferguson, B., Hernandez, R. D., Hirani, K., Kehrer-Sawatzki, H., Kolb, J., Patil, S., Pu, L.-L., Ren, Y., Smith, D. G., Wheeler, D. A., Schenck, I., Ball, E. V., Chen, R., Cooper, D. N., Giardine, B., Hsu, F., Kent, W. J., Lesk, A., Nelson, D. L., O'brien, W. E., Prufer, K., Stenson, P. D., Wallace, J. C., Ke, H., Liu, X.-M., Wang, P., Xiang, A. P., Yang, F., Barber, G. P., Haussler, D., Karolchik, D., Kern, A. D., Kuhn, R. M., Smith, K. E. and Zwieg, A. S. (2007), 'Evolutionary and biomedical insights from the rhesus macaque genome.', *Science* **316**(5822), 222–234.

Giudicelli, M. D., Serazin, V., Le Sciellour, C. R., Albert, M., Selva, J. and Giudicelli, Y. (2008), 'Increased achondroplasia mutation frequency with advanced age and evidence for G1138A mosaicism in human testis biopsies.', *Fertil Steril* **89**(6), 1651–1656.

Glaser, R. L., Jiang, W., Boyadjiev, S. A., Tran, A. K., Zachary, A. A., Van Maldergem, L., Johnson, D., Walsh, S., Oldridge, M., Wall, S. A., Wilkie, A. O. and Jabs, E. W. (2000), 'Paternal origin of FGFR2 mutations in sporadic cases of Crouzon syndrome and Pfeiffer syndrome.', *Am J Hum Genet* **66**(3), 768–777.

Goetting-Minesky, M. P. and Makova, K. D. (2006), 'Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates.', *J Mol Evol* **63**(4), 537–44.

Goldman, N. and Yang, Z. (1994), 'A codon-based model of nucleotide substitution for protein-coding dna sequences.', *Mol Biol Evol* **11**(5), 725–36.

Gong, F., Kwon, Y. and Smerdon, M. J. (2005), 'Nucleotide excision repair in chromatin and the right of entry.', *DNA Repair (Amst)* **4**(8), 884–96.

Goriely, A., McVean, G. A. T., Rojmyr, M., Ingemarsson, B. and Wilkie, A. O. M. (2003), 'Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line.', *Science* **301**(5633), 643–6.

Green, P., Ewing, B., Miller, W., Thomas, P. J. and Green, E. D. (2003), 'Transcription-associated mutational asymmetry in mammalian evolution.', *Nat Genet* **33**(4), 514–7.

Grunewald, S., Paasch, U., H.-J, G. and U, A. (2005), 'Mature human spermatozoa do not transcribe novel RNA', *Andrologia* **37**(2-3), 69–71.

Hanawalt, P. C. and Spivak, G. (2008), 'Transcription-coupled DNA repair: two decades of progress and surprises.', *Nat Rev Mol Cell Biol* **9**(12), 958–970.

Hedrick, P. W. (2007), 'Sex: differences in mutation, recombination, selection, gene flow, and genetic drift.', *Evolution* **61**(12), 2750–2771.

Hendriks, G., Calleja, F., Vrieling, H., Mullenders, L. H. F., Jansen, J. G. and de Wind, N. (2008), 'Gene transcription increases DNA damage-induced mutagenesis in mammalian stem cells.', *DNA Repair (Amst)* **7**(8), 1330–1339.

Hernandez, R. D., Williamson, S. H., Zhu, L. and Bustamante, C. D. (2007), 'Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans.', *Mol Biol Evol* **24**(10), 2196–2202.

Hernandez, R., Williamson, S. and Bustamante, C. (2007), 'Context dependence, ancestral misidentification, and spurious signatures of natural selection.', *Mol Biol Evol* .

Hess, S. T., Blake, J. D. and Blake, R. D. (1994), 'Wide variations in neighbor-dependent substitution rates.', *J Mol Biol* **236**(4), 1022–33.

Hill, K. A., Halangoda, A., Heinmoeller, P. W., Gonzalez, K., Chitaphan, C., Longmate, J., Scaringe, W. A., Wang, J.-C. and Sommer, S. S. (2005), 'Tissue-specific time courses of spontaneous mutation frequency and deviations in mutation pattern are observed in middle to late adulthood in Big Blue mice.', *Environ Mol Mutagen* **45**(5), 442–454.

Hobolth, A., Christensen, O. F., Mailund, T. and Schierup, M. H. (2007), 'Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model.', *PLoS Genet* **3**(2), e7.

Hurst, L. D. and Ellegren, H. (1998), 'Sex biases in the mutation rate.', *Trends Genet* **14**(11), 446–52.

Hurst, L. D. and Williams, E. J. (2000), 'Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores.', *Gene* **261**(1), 107–14.

Huttley, G. A. (2004), 'Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals.', *Mol Biol Evol* **21**(9), 1760–8.

Huttley, G. A., Jakobsen, I. B., Wilson, S. R. and Easteal, S. (2000), 'How important is DNA replication for mutagenesis?', *Mol Biol Evol* **17**(6), 929–37.

Huttley, G., Wakefield, M. and Easteal, S. (2007), 'Rates of genome evolution and branching order from whole genome analysis.', *Mol Biol Evol* .

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 U S A* **101**(39), 13994–4001.

Jansen, J., Olsen, A. K., Wiger, R., Naegeli, H., de Boer, P., van Der Hoeven, F., Holme, J. A., Brunborg, G. and Mullenders, L. (2001), 'Nucleotide excision repair in rat male germ cells: low level of repair in intact cells contrasts with high dual incision activity in vitro.', *Nucleic Acids Res* **29**(8), 1791–800.

Jensen, J. and Pedersen, A. (2000), 'Probabilistic models of DNA sequence evolution with context dependent rates of substitution', *Advances In Applied Probability* **32**, 499–517.

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., 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**(4), 528–38.

Jensen, T. K., Carlsen, E., Jorgensen, N., Berthelsen, J. G., Keiding, N., Christensen, K., Petersen, J. H., Knudsen, L. B. and Skakkebaek, N. E. (2002), 'Poor semen quality may contribute to recent decline in fertility rates.', *Hum Reprod* **17**(6), 1437–1440.

Ji, B. T., Shu, X. O., Linet, M. S., Zheng, W., Wacholder, S., Gao, Y. T., Ying, D. M. and Jin, F. (1997), 'Paternal cigarette smoking and the risk of childhood cancer among offspring of nonsmoking mothers.', *J Natl Cancer Inst* **89**(3), 238–244.

Johnson, J., Canning, J., Kaneko, T., Pru, J. K. and Tilly, J. L. (2004), 'Germline stem cells and follicular renewal in the postnatal mammalian ovary.', *Nature* **428**(6979), 145–150.

Jung, A., Schuppe, H.-C. and Schill, W.-B. (2003), 'Are children of older fathers at risk for genetic disorders?', *Andrologia* **35**(4), 191–199.

Kaessmann, H., Wiebe, V., Weiss, G. and Paabo, S. (2001), 'Great ape dna sequences reveal a reduced diversity and an expansion in humans.', *Nat Genet* **27**(2), 155–156.

Kahn, N. W. and Quinn, T. W. (1999), 'Male-driven evolution among Eoaves? a test of the replicative division hypothesis in a heterogametic female (ZW) system.', *J Mol Evol* **49**(6), 750–9.

Ketterling, R. P., Vielhaber, E., Bottema, C. D., Schaid, D. J., Cohen, M. P., Sexauer, C. L. and Sommer, S. S. (1993), 'Germ-line origins of mutation in families with hemophilia B: the sex ratio varies with the type of mutation.', *Am J Hum Genet* **52**(1), 152–166.

Khil, P. P., Smirnova, N. A., Romanienko, P. J. and Camerini-Otero, R. D. (2004), 'The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation.', *Nat Genet* **36**(6), 642–646.

Kim, N., Abdulovic, A. L., Gealy, R., Lippert, M. J. and Jinks-Robertson, S. (2007), 'Transcription-associated mutagenesis in yeast is directly proportional to the level of gene expression and influenced by the direction of DNA replication.', *DNA Repair (Amst)* **6**(9), 1285–1296.

Kim, S.-H., Elango, N., Warden, C., Vigoda, E. and Yi, S. V. (2006), 'Heterogeneous genomic molecular clocks in primates.', *PLoS Genet* **2**(10), e163.

Kimura, M. (1983), *The neutral theory of molecular evolution*, Cambridge University Press, Cambridge.
URL: *http://www.loc.gov/catdir/description/cam041/82022225.html*

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J. G., Easton, B. C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., Lozupone, C., McDonald, D., Robeson, M., Sammut, R., Smit, S., Wakefield, M. J., Widmann, J., Wikman, S., Wilson, S., Ying, H. and Huttley, G. A. (2007), 'PyCogent: a toolkit for making sense from sequence.', *Genome Biol* **8**(8), R171.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R. and Stefansson, K. (2002), 'A high-resolution recombination map of the human genome.', *Nat Genet* **31**(3), 241–7.

Krawczak, M., Ball, E. V. and Cooper, D. N. (1998), 'Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes.', *Am J Hum Genet* **63**(2), 474–88.

Lander, E. S. e. a. (2001), 'Initial sequencing and analysis of the human genome.', *Nature* **409**(6822), 860–921.

Lawson, L.-J. and Hewitt, G. M. (2002), 'Comparison of substitution rates in ZFX and ZFY introns of sheep and goat related species supports the hypothesis of male-biased mutation rates.', *J Mol Evol* **54**(1), 54–61.

Lazaro, C., Gaona, A., Ainsworth, P., Tenconi, R., Vidaud, D., Kruyer, H., Ars, E., Volpini, V. and Estivill, X. (1996), 'Sex differences in mutational rate and mutational mechanism in the NF1 gene in neurofibromatosis type 1 patients.', *Hum Genet* **98**(6), 696–699.

Leadon, S. A. and Lawrence, D. A. (1991), 'Preferential repair of DNA damage on the transcribed strand of the human metallothionein genes requires RNA polymerase II.', *Mutat Res* **255**(1), 67–78.

Lercher, M. J., Williams, E. J. 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**(11), 2032–9.

Lewontin, R. C. (1989), 'Inferring the number of evolutionary events from dna coding sequence differences.', *Mol Biol Evol* **6**(1), 15–32.

Li, R., Johnson, A. B., Salomons, G. S., van der Knaap, M. S., Rodriguez, D., Boespflug-Tanguy, O., Gorospe, J. R., Goldman, J. E., Messing, A. and Brenner, M. (2006), 'Propensity for paternal inheritance of de novo mutations in Alexander disease.', *Hum Genet* **119**(1-2), 137–144.

Li, S. and Smerdon, M. J. (2004), 'Dissecting transcription-coupled and global genomic repair in the chromatin of yeast GAL1-10 genes.', *J Biol Chem* **279**(14), 14418–26.

Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H. and Hewett-Emmett, D. (1996), 'Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis.', *Mol Phylogenet Evol* **5**(1), 182–7.

Lieb, M. and Rehmat, S. (1997), '5-methylcytosine is not a mutation hot spot in nondividing escherichia coli.', *Proc Natl Acad Sci U S A* **94**(3), 940–945.

Lindahl, T. and Nyberg, B. (1974), 'Heat-induced deamination of cytosine residues in deoxyribonucleic acid.', *Biochemistry* **13**(16), 3405–3410.

Lindsay, H., Yap, V. B., Ying, H. and Huttley, G. A. (2008), 'Pitfalls of the most commonly used models of context dependent substitution.', *Biol Direct* **3**, 52.

Lippert, M. J., Chen, Q. and Liber, H. L. (1998), 'Increased transcription decreases the spontaneous mutation rate at the thymidine kinase locus in human cells.', *Mutat Res* **401**(1-2), 1–10.

Lucifero, D., Mertineit, C., Clarke, H. J., Bestor, T. H. and Trasler, J. M. (2002), 'Methylation dynamics of imprinted genes in mouse germ cells.', *Genomics* **79**(4), 530–538.

Majewski, J. (2003), 'Dependence of mutational asymmetry on gene-expression levels in the human genome.', *Am J Hum Genet* **73**(3), 688–92.

Makova, K. D. and Li, W.-H. (2002), 'Strong male-driven evolution of DNA sequences in humans and apes.', *Nature* **416**(6881), 624–6.

Malaspina, D., Corcoran, C., Fahim, C., Berman, A., Harkavy-Friedman, J., Yale, S., Goetz, D., Goetz, R., Harlap, S. and Gorman, J. (2002), 'Paternal age and sporadic schizophrenia: evidence for de novo mutations.', *Am J Med Genet* **114**(3), 299–303.

Marchetti, F. and Wyrobek, A. J. (2008), 'DNA repair decline during mouse spermiogenesis results in the accumulation of heritable DNA damage.', *DNA Repair (Amst)* **7**(4), 572–581.

Martin, A. P. and Palumbi, S. R. (1993), 'Body size, metabolic rate, generation time, and the molecular clock.', *Proc Natl Acad Sci U S A* **90**(9), 4087–4091.

Martomo, S. A. and Mathews, C. K. (2002), 'Effects of biological DNA precursor pool asymmetry upon accuracy of DNA replication in vitro.', *Mutat Res* **499**(2), 197–211.

McVean, G. (2000), 'Evolutionary genetics: what is driving male mutation?', *Curr Biol* **10**(22), R834–5.

McVean, G. T. and Hurst, L. D. (1997), 'Evidence for a selectively favourable reduction in the mutation rate of the X chromosome.', *Nature* **386**(6623), 388–92.

Mimault, C., Giraud, G., Courtois, V., Cailloux, F., Boire, J. Y., Dastugue, B. and Boespflug-Tanguy, O. (1999), 'Proteolipoprotein gene analysis in 82 patients with sporadic Pelizaeus-Merzbacher Disease: duplications, the major cause of the disease, originate more frequently in male germ cells, but point mutations do not. the clinical european network on brain dysmyelinating disease.', *Am J Hum Genet* **65**(2), 360–369.

Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. and Yasunaga, T. (1987), 'Male-driven molecular evolution: a model and nucleotide sequence analysis.', *Cold Spring Harb Symp Quant Biol* **52**, 863–7.

Moloney, D. M., Slaney, S. F., Oldridge, M., Wall, S. A., Sahlin, P., Stenman, G. and Wilkie, A. O. (1996), 'Exclusive paternal origin of new mutations in Apert syndrome.', *Nat Genet* **13**(1), 48–53.

Monk, M., Boubelik, M. and Lehnert, S. (1987), 'Temporal and regional changes in dna methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development.', *Development* **99**(3), 371–382.

Morgan, A. R. (1993), 'Base mismatches and mutagenesis: how important is tautomerism?', *Trends Biochem Sci* **18**(5), 160–163.

Morgan, H. D., Santos, F., Green, K., Dean, W. and Reik, W. (2005), 'Epigenetic reprogramming in mammals.', *Hum Mol Genet* **14 Spec No 1**, R47–58.

Morison, I. M., Paton, C. J. and Cleverley, S. D. (2001), 'The imprinted gene and parent-of-origin effect database.', *Nucleic Acids Res* **29**(1), 275–276.

Morton, B. R., Bi, I. V., McMullen, M. D. and Gaut, B. S. (2006), 'Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition.', *Genetics* **172**(1), 569–77.

Morton, B. R., Dar, V.-u.-N. and Wright, S. I. (2009), 'Analysis of site frequency spectra from arabidopsis with context-dependent corrections for ancestral misinference.', *Plant Physiol* **149**(2), 616–624.

Muse, S. V. (1996), 'Estimating synonymous and nonsynonymous substitution rates.', *Mol Biol Evol* **13**(1), 105–114.

Muse, S. V. and Gaut, B. S. (1994), 'A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.', *Mol Biol Evol* **11**(5), 715–724.

Nachman, M. W. and Crowell, S. L. (2000), 'Estimate of the mutation rate per nucleotide in humans.', *Genetics* **156**(1), 297–304.

Ochman, H. (2003), 'Neutral mutations and neutral substitutions in bacterial genomes.', *Mol Biol Evol* **20**(12), 2091–2096.

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y. and Okada, N. (2003), 'Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates.', *Genome Biol* **4**(11), R74.

Olsen, A.-K., Lindeman, B., Wiger, R., Duale, N. and Brunborg, G. (2005), 'How do male germ cells handle dna damage?', *Toxicol Appl Pharmacol* **207**(2 Suppl), 521–531.

Olshan, A. F., Schnitzer, P. G. and Baird, P. A. (1994), 'Paternal age and the risk of congenital heart defects.', *Teratology* **50**(1), 80–84.

Ono, T., Ikehata, H., Nakamura, S., Saito, Y., Hosoi, Y., Takai, Y., Yamada, S., Onodera, J. and Yamamoto, K. (2000), 'Age-associated increase of spontaneous mutant frequency and molecular nature of mutation in newborn and old lacZ-transgenic mouse.', *Mutat Res* **447**(2), 165–177.

Oscamou, M., McDonald, D., Yap, V. B., Huttley, G. A., Lladser, M. E. and Knight, R. (2008), 'Comparison of methods for estimating the nucleotide substitution matrix.', *BMC Bioinformatics* **9**, 511.

Pacholczyk, M. and Kimmel, M. (2005), 'Analysis of differences in amino acid substitution patterns, using multilevel G-tests.', *C R Biol* **328**(7), 632–641.

Parapanov, R., Nussle, S. and Vogel, P. (2007), 'Cycle length of spermatogenesis in shrews (mammalia: soricidae) with high and low metabolic rates and different mating systems.', *Biol Reprod* **76**(5), 833–840.

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. and Reich, D. (2006), 'Genetic evidence for complex speciation of humans and chimpanzees.', *Nature* **441**(7097), 1103–1108.

Pecon Slattery, J. and O'Brien, S. J. (1998), 'Patterns of Y and X chromosome DNA sequence divergence during the Felidae radiation.', *Genetics* **148**(3), 1245–1255.

Pedersen, A. M. and Jensen, J. L. (2001), 'A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames.', *Mol Biol Evol* **18**(5), 763–76.

Penrose, L. S. (1957), 'Parental age in achondroplasia and mongolism', *Am J Hum Genet* **9**(3), 167–9.

Pepling, M. E. (2006), 'From primordial germ cell to primordial follicle: mammalian female germ cell development.', *Genesis* **44**(12), 622–632.

Pfeifer, G. P. (2000), 'p53 mutational spectra and the role of methylated CpG sequences.', *Mutat Res* **450**(1-2), 155–166.

Picton, H., Briggs, D. and Gosden, R. (1998), 'The molecular basis of oocyte growth and development.', *Mol Cell Endocrinol* **145**(1-2), 27–37.

Pink, C. and Hurst, L. (2009), 'Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents.', *Mol Biol Evol* .

Pink, C. J., Swaminathan, S. K., Dunham, I., Rogers, J., Ward, A. and Hurst, L. D. (2009), 'Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates', *Genome Biol Evol* **Advanced access April 30**.

Prendergast, J. G. D., Campbell, H., Gilbert, N., Dunlop, M. G., Bickmore, W. A. and Semple, C. A. M. (2007), 'Chromatin structure and evolution in the human genome.', *BMC Evol Biol* **7**, 72.

Presgraves, D. C. and Yi, S. V. (2009), 'Doubts about complex speciation between humans and chimpanzees.', *Trends Ecol Evol* **24**(10), 533–540.

Qin, J., Calabrese, P., Tiemann-Boege, I., Shinde, D. N., Yoon, S.-R., Gelfand, D., Bauer, K. and Arnheim, N. (2007), 'The molecular anatomy of spontaneous germline mutations in human testes.', *PLoS Biol* **5**(9), e224.

Radford, I. R. and Lobachevsky, P. N. (2008), 'Clustered DNA lesion sites as a source of mutations during human colorectal tumourigenesis.', *Mutat Res* **646**(1-2), 60–68.

Rannan-Eliya, S. V., Taylor, I. B., De Heer, I. M., Van Den Ouweland, A. M. W., Wall, S. A. and Wilkie, A. O. M. (2004), 'Paternal origin of FGFR3 mutations in Muenke-type craniosynostosis.', *Hum Genet* **115**(3), 200–207.

Reichenberg, A., Gross, R., Weiser, M., Bresnahan, M., Silverman, J., Harlap, S., Rabinowitz, J., Shulman, C., Malaspina, D., Lubin, G., Knobler, H. Y., Davidson, M. and Susser, E. (2006), 'Advancing paternal age and autism.', *Arch Gen Psychiatry* **63**(9), 1026–1032.

Richards, F. M., Payne, S. J., Zbar, B., Affara, N. A., Ferguson-Smith, M. A. and Maher, E. R. (1995), 'Molecular analysis of de novo germline mutations in the von Hippel-Lindau disease gene.', *Hum Mol Genet* **4**(11), 2139–2143.

Risch, N., Reich, E. W., Wishnick, M. M. and McCarthy, J. G. (1987), 'Spontaneous mutation and parental age in humans.', *Am J Hum Genet* **41**(2), 218–248.

Rogozin, I. B., Malyarchuk, B. A., Pavlov, Y. I. and Milanesi, L. (2005), 'From context-dependence of mutations to molecular mechanisms of mutagenesis.', *Pac Symp Biocomput* pp. 409–20.

Sandstedt, S. A. and Tucker, P. K. (2005), 'Male-driven evolution in closely related species of the mouse genus *Mus*.', *J Mol Evol* **61**(1), 138–144.

Saunders, C. T. and Green, P. (2007), 'Insights from modeling protein evolution with context-dependent mutation and asymmetric amino acid selection.', *Mol Biol Evol* **24**(12), 2632–2647.

Schmegner, C., Hameister, H., Vogel, W. and Assum, G. (2007), 'Isochores and replication time zones: a perfect match.', *Cytogenet Genome Res* **116**(3), 167–172.

Schoniger, M. and von Haeseler, A. (1994), 'A stochastic model for the evolution of autocorrelated dna sequences.', *Mol Phylogenet Evol* **3**(3), 240–247.

Schranz, H. W., Yap, V. B., Easteal, S., Knight, R. and Huttley, G. A. (2008), 'Pathological rate matrices: from primates to pathogens.', *BMC Bioinformatics* **9**, 550.

Schultz, N., Hamra, F. K. and Garbers, D. L. (2003), 'A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets.', *Proc Natl Acad Sci U S A* **100**(21), 12201–6.

Shen, J. C., Rideout, W. M. r. and Jones, P. A. (1994), 'The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA.', *Nucleic Acids Res* **22**(6), 972–976.

Shimmin, L. C., Chang, B. H., Hewett-Emmett, D. and Li, W. H. (1993), 'Potential problems in estimating the male-to-female mutation rate ratio from dna sequence data.', *J Mol Evol* **37**(2), 160–6.

Siepel, A. and Haussler, D. (2004), 'Phylogenetic estimation of context-dependent substitution rates by maximum likelihood.', *Mol Biol Evol* **21**(3), 468–88.

Sigurdsson, M. I., Smith, A. V., Bjornsson, H. T. and Jonsson, J. J. (2009), 'HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination.', *Genome Res* **19**(4), 581–589.

Singh, N. D., Arndt, P. F. and Petrov, D. A. (2006), 'Minor shift in background substitutional patterns in the *Drosophila saltans* and *willistoni* lineages is insufficient to explain GC content of coding sequences.', *BMC Biol* **4**, 37.

Smit, A. F. (1999), 'Interspersed repeats and other mementos of transposable elements in mammalian genomes.', *Curr Opin Genet Dev* **9**(6), 657–663.

Smith, N. G. C., Webster, M. T. and Ellegren, H. (2002), 'Deterministic mutation rate variation in the human genome.', *Genome Res* **12**(9), 1350–6.

Smith, N. G. C., Webster, M. T. and Ellegren, H. (2003), 'A low rate of simultaneous double-nucleotide mutations in primates.', *Mol Biol Evol* **20**(1), 47–53.

Smith, N. G. and Hurst, L. D. (1999), 'The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate?', *Genetics* **152**(2), 661–73.

Splendore, A., Jabs, E. W., Felix, T. M. and Passos-Bueno, M. R. (2003), 'Parental origin of mutations in sporadic cases of Treacher Collins syndrome.', *Eur J Hum Genet* **11**(9), 718–722.

Stancheva, I. (2005), 'Caught in conspiracy: cooperation between DNA methylation and histone H3K9 methylation in the establishment and maintenance of heterochromatin.', *Biochem Cell Biol* **83**(3), 385–95.

Stoltzfus, A. (2008), 'Evidence for a predominant role of oxidative damage in germline mutation in mammals.', *Mutat Res* **644**(1-2), 71–73.

Sueoka, N. (1995), 'Intrastrand parity rules of DNA base composition and usage biases of synonymous codons.', *J Mol Evol* **40**(3), 318–25.

Swales, A. K. E. and Spears, N. (2005), 'Genomic imprinting and reproduction.', *Reproduction* **130**(4), 389–399.

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**(3), 565–73.

Tiemann-Boege, I., Navidi, W., Grewal, R., Cohn, D., Eskenazi, B., Wyrobek, A. J. and Arnheim, N. (2002), 'The observed human sperm mutation frequency cannot explain the achondroplasia paternal age effect.', *Proc Natl Acad Sci U S A* **99**(23), 14952–7.

Topal, M. D. and Fresco, J. R. (1976), 'Complementary base pairing and the origin of substitution mutations.', *Nature* **263**(5575), 285–289.

Touchon, M., Nicolay, S., Audit, B., Brodie of Brodie, E.-B., d'Aubenton Carafa, Y., Arneodo, A. and Thermes, C. (2005), 'Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins.', *Proc Natl Acad Sci U S A* **102**(28), 9836–41.

Trappe, R., Laccone, F., Cobilanschi, J., Meins, M., Huppke, P., Hanefeld, F. and Engel, W. (2001), 'MECP2 mutations in sporadic cases of Rett syndrome are almost exclusively of paternal origin.', *Am J Hum Genet* **68**(5), 1093–1101.

Trasler, J. M. (2006), 'Gamete imprinting: setting epigenetic patterns for the next generation.', *Reprod Fertil Dev* **18**(1-2), 63–9.

Trasler, J. M. (2009), 'Epigenetics in spermatogenesis.', *Mol Cell Endocrinol* **306**(1-2), 33–36.

Turner, J. M. A. (2007), 'Meiotic sex chromosome inactivation.', *Development* **134**(10), 1823–1831.

van der Heijden, G. W., Ramos, L., Baart, E. B., van den Berg, I. M., Derijck, A. A. H. A., van der Vlag, J., Martini, E. and de Boer, P. (2008), 'Sperm-derived histones contribute to zygotic chromatin in humans.', *BMC Dev Biol* **8**, 34.

Wakeley, J. (2008), 'Complex speciation of humans and chimpanzees.', *Nature* **452**(7184), E3–4; discussion E4.

Walter, C. A., Intano, G. W., McMahan, C. A., Kelner, K., McCarrey, J. R. and Walter, R. B. (2004), 'Mutation spectral changes in spermatogenic cells obtained from old mice.', *DNA Repair (Amst)* **3**(5), 495–504.

Wang, D., Kreutzer, D. A. and Essigmann, J. M. (1998), 'Mutagenicity and repair of oxidative DNA damage: insights from studies using defined lesions.', *Mutat Res* **400**(1-2), 99–115.

Watson, J. D. and Crick, F. H. (1953), 'The structure of DNA.', *Cold Spring Harb Symp Quant Biol* **18**, 123–131.

Webster, M. T., Smith, N. G. C. and Ellegren, H. (2003), 'Compositional evolution of noncoding DNA in the human and chimpanzee genomes.', *Mol Biol Evol* **20**(2), 278–86.

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**(6), 1468–74.

Weiss, G. and von Haeseler, A. (2003), 'Testing substitution models within a phylogenetic tree.', *Mol Biol Evol* **20**(4), 572–578.

Whelan, S. (2008), 'Spatial and temporal heterogeneity in nucleotide sequence evolution.', *Mol Biol Evol* **25**(8), 1683–1694.

Whelan, S. and Goldman, N. (2004), 'Estimating the frequency of events that cause multiple-nucleotide changes.', *Genetics* **167**(4), 2027–43.

Wilkin, D. J., Szabo, J. K., Cameron, R., Henderson, S., Bellus, G. A., Mack, M. L., Kaitila, I., Loughlin, J., Munnich, A., Sykes, B., Bonaventure, J. and Francomano, C. A. (1998), 'Mutations in fibroblast growth-factor receptor 3 in sporadic cases of achondroplasia occur exclusively on the paternally derived chromosome.', *Am J Hum Genet* **63**(3), 711–716.

Williams, S. K. and Tyler, J. K. (2007), 'Transcriptional regulation by chromatin disassembly and reassembly.', *Curr Opin Genet Dev* **17**(2), 88–93.

Wolfe, K. H., Sharp, P. M. and Li, W. H. (1989), 'Mutation rates differ among regions of the mammalian genome.', *Nature* **337**(6204), 283–5.

Woodfine, K., Fiegler, H., Beare, D. M., Collins, J. E., McCann, O. T., Young, B. D., Debernardi, S., Mott, R., Dunham, I. and Carter, N. P. (2004), 'Replication timing of the human genome.', *Hum Mol Genet* **13**(2), 191–202.

Workman, J. L. and Kingston, R. E. (1998), 'Alteration of nucleosome structure as a mechanism of transcriptional regulation.', *Annu Rev Biochem* **67**, 545–579.

Wrobel, G. and Primig, M. (2005), 'Mammalian male germ cells are fertile ground for expression profiling of sexual reproduction.', *Reproduction* **129**(1), 1–7.

Xu, G., Spivak, G., Mitchell, D. L., Mori, T., McCarrey, J. R., McMahan, C. A., Walter, R. B., Hanawalt, P. C. and Walter, C. A. (2005), 'Nucleotide excision repair activity varies among murine spermatogenic cell types.', *Biol Reprod* **73**(1), 123–130.

Yap, V. B., Lindsay, H., Easteal, S. and Huttley, G. (2010), 'Estimates of the effect of natural selection on protein-coding content.', *Mol Biol Evol* **27**(3), 726–734.

Ying, H., Epps, J., Williams, R. and Huttley, G. (2010), 'Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair.', *Mol Biol Evol* **27**(3), 637–649.

Yoon, S.-R., Qin, J., Glaser, R. L., Wang Jabs, E., Wexler, N. S., Sokol, R., Arnheim, N. and Calabrese, P. (2009), 'The ups and downs of mutation frequencies during aging can account for the Apert syndrome paternal age effect.', *PLoS Genet* **5**(7), e1000558.

Zhao, G. Q. and Garbers, D. L. (2002), 'Male germ cell specification and differentiation.', *Dev Cell* **2**(5), 537–547.

Zheng, T., Ichiba, T. and Morton, B. R. (2007), 'Assessing substitution variation across sites in grass chloroplast DNA.', *J Mol Evol* **64**(6), 605–613.