Inferences On Selection And Mutation From Substitution Rate Differences In Both Recombining And Non-Recombining Regions Of Sex Chromosomes

by

David Thomas Gerrard

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
The University of Birmingham
2004

Synopsis

This thesis comprises an introductory chapter and four experimental chapters. The introduction is a literature review of the forms and evolution of sex chromosomes and of the molecular evolutionary processes that affect sex-linked genes. Special attention is paid to the commonly and convergently evolved trait of the degenerative Y, whereby these non-recombining chromosomes progressively lose functional genetic material over time. The differences in inheritance, in physical structure and in genetic composition that exist between the sex chromosomes and autosomes, between the X and the Y and even along a single sex chromosome may allow us to partition and contrast the effects of the many hypothesised mechanisms that govern the evolution of entire genomes.

The experimental chapters are presented in chronological order but also approach a finer scale; from the generation of a genomic distribution, through an investigation of a class of genes down to detailed analyses of specific sex linked genes.

Chapter II deals with sequences in the pseudo-autosomal region (PAR) of the mammalian sex chromosomes, the only region in which the X and Y chromosomes are able to recombine. Due to the small size of the PAR, the recombination rate per nucleotide is higher than any other comparable sized region in the genome. A link between recombination and mutation, suggested by several other studies, is investigated using primate sequence. The chapter describes how 51 pairs of human and orangutan DNA sequences were aligned to generate a distribution of divergence values between these two species. This distribution was then used to show that silent sites in the PAR are evolving at an abnormally high rate; consistent with the hypothesis that recombination is mutagenic. A paper published in the journal ***Gene*** and using this result is included after the chapter.

In contrast to the previous chapter, Chapter III investigates three pairs of X and Y linked genes that have not recombined with each other for tens of millions of years. By sequencing and aligning these genes in a range of primates, I was able to test for differences in the rate of adaptive evolution between different sites along the sequence. As well as a reduced level of selective constraint on the Y chromosome relative to the X, I found evidence that positive selection is still able to drive the evolution of genes in a non-recombining Y chromosome after such a long time. This result is surprising and contradictory to several predictions of the demise of Y chromosomes. Chapter III has been submitted to the journal ***Molecular Biology and Evolution***.

In Chapter IV these methods are applied to Amelogenin (AMEL), the gene that helps to lay down tooth enamel. This gene used to lie inside the sex chromosome PAR but since monkeys diverged from lemurs, it has existed as X and Y forms in the lineage leading to humans. In attempting to discern if the Y linked amelogenin is still under selection or evolving neutrally, several conflicting patterns were discovered. Whilst some of the primate AMEL sequences appeared to be recombining, it was found to be more likely that recurrent mutation at specific sites in an otherwise conserved sequence were generating a false signal. Y linked amelogenin is still conserved, but is adjusting to life on a non-recombining chromosome with its associated mutation bias and reduction in constraint.

The final chapter changes kingdoms to the newly evolved plant sex chromosomes in Silene (Campion). In this new system it appears that the Y chromosome has not yet come to terms with life on its own and the X chromosome cannot chose which 'species' it belongs to. In order to determine if X chromosomes are regularly passing between the species *S. dioica* and *S. latifolia* a section of the gene SlX1 was sequenced in a sample of European plants. The relationships of the SlX1 sequences from X and Y chromosomes were found to be more complicated than previously thought and it is possible that recombination between X and Y ceased recently but independently in these two species.

To **Claire**,

"you're the reason I am leaving
to work in another country"

Work is the grand cure of all the
maladies and miseries that
ever beset mankind

*--**Thomas Carlyle***
(1795-1881)

Acknowledgements

Primarily I have to thank Dmitry for his patience and honesty and for taking me on as his first PhD student; a debt I am unlikely ever to be able to repay.

I should also thank the BBSRC for food, lodgings and the occasional pint of beer.

Though not at Birmingham any more, I would like to thank, Drs Joe Ironside and Ravinder Kanda for their help and guidance in the lab and their interpretation of life and genetics. Thanks also to the various people who should have come more often to Journal Club to stop Dmitry choosing such long papers.

Thank you to the constantly regenerating members of Room S208, there are armies that don't eat as much as you lot, but they probably don't laugh as much either. Keep it up.

Thanks also to all the good friends I have made whilst in Birmingham. This means the Wayfarers, from which I have drawn housemates, friends-for-big-adventures, opportunities for fresh air and a girlfriend.

And finally my mum, my brother and my sisters, thank you for not seeming to mind that you barely saw anything of me for the last three years. I don't know how you coped!

Table of contents

List of illustrations

List of tables

List of definitions & abbreviations

Aneuploid	A karyotype with an atypical complement of chromosomes e.g. the 47, XXY sex karyotype
Anthropoidea	The suborder including Monkeys from both the New World and the Old World (including apes).
Autosome	A non-sex-linked nuclear chromosome.
Catarrhini	The family of Old world monkeys including Apes.
cM	centiMorgan
Gametologue	One of a pair of genes present on X and Y chromosomes descended from alleles of a common ancestral gene and diverged since the cessation of recombination in their genetic locality.
Gynodioecious	Having hermaphroditic and female individuals in the same population.
Heterogametic	Of a karyotype, having two different sex chromosomes e.g. XY (males) or ZW (females) c.f. Homogametic
Homogametic	Of a karyotype, having two equal sex chromosomes e.g. XX (females) or ZZ (males) c.f. Heterogametic
Indel	An insertion or deletion mutation in nucleotide sequence. When the ancestral state is not known, gaps in a sequence alignment could equally be insertions or deletions.
K_a	The rate of non-synonymous substitution per non-synonymous site.
K_s	The rate of synonymous substitution per synonymous site.
LRT	Likelihood Ratio Test, used to test if one model better describes the data than another.
Mb	Megabase: a physical chromosome distance of one million consecutive nucleotides.
MYA	Million Years Ago
N_e	The effective population size. The size of theoretical ideal population that would be required

	to explain the levels of diversity seen in the observed population.
NRY	Non-Recombining-Y. That part of the Y chromosome which does not recombine with the X chromosome during meiosis. In mammals, everything but the pseudoautosomal region (PAR).
NW	New World. Referring to the monkey species of South America.
OW	Old World. Referring to the monkey (and ape) species of Africa and Asia.
PAR	Pseudoautosomal region
PAB	Pseudoautosomal boundary.
Platyrrhini	The family of New World monkeys.
Pseudoautosomal	A region of homology between sex chromosomes which may recombine during meiosis so that alleles in this region segregate as would those on the non-sex chromosomes (autosomes).
Strepsirrhini	The family of Lemurs
X-NRY	Those X linked genes which appear to have functional Y linked homologues located within the non-recombining region (NRY) of the Y. Usually both genes are present in only one copy.
ω	The symbol used to represent the ratio of Ka to Ks in the programme PAML (Yang 1997). It may be used for a specific lineage within a phylogenetic tree or for some sites within a sequence.

I AN INTRODUCTION TO SEX CHROMOSOMES AND THE PROCESSES GOVERNING THEIR MOLECULAR EVOLUTION

Dave T. Gerrard

School of Biosciences,

The University of Birmingham,

Edgbaston, Birmingham, B15 2TT, UK

I.1 Sex Chromosomes

I.1.1 Anatomy of the modern mammalian Y chromosome - e.g. mine

My Y chromosome, and I'm fairly sure I have one because I seem to have accidentally PCR amplified it a couple of times, is a typical modern mammalian Y chromosome. In part, it is orthologous to all the Y chromosomes in all the male mammals of the world, and related through descent all the way back to the granddad of all mammals, in which the first Y chromosome was created from a normal chromosome (an autosome) (Burgoyne 1982; Lahn and Page 1999). Interestingly, my Y chromosome shares physical and genetical properties with the Y chromosomes in other, more distant taxa; even though they are not related, they have evolved independently yet convergently (Charlesworth 1996).

Very simply, when Y chromosomes evolve to reside in just one sex, they lose the ability to exchange genetic material by recombination with their old partner, the X. Whilst genes on a bachelor Y chromosome are free to evolve male specific functions and not worry about life as a female, they lose access to the conjugal rights of recombination, a fundamental genetic process which keeps genes healthy and able to adapt (Felsenstein 1974). The result for the Y chromosome is degeneration: the accumulation of harmful mutations and the loss of most of the genes it once called its own (Rice 1994) (Charlesworth and Charlesworth 2000).

I also have an X chromosome, everyone does, around half the population have two. Whilst the Y confines itself to males, the X leads a double life, being present in both sexes. Genes on the X chromosome are still able to recombine and stay healthy, but must also adapt to cope with an imbalance in copy number between males (one X) and females (two Xs).

Genetic forces such as selection, mutation and drift are experienced differently by genes on the X and Y chromosomes and this has led to great interest in their molecular evolution. The contrast in genetic content between X and Y and between them and the autosomes may help to reveal how important such forces are in the evolution of the entire genome and, in fact, all genomes.

In this thesis I hope to present some of the differing forms of molecular evolution occurring on sex chromosomes and link the observations with wider processes. I will begin by reviewing the diversity in form and evolution of sex chromosomes and the processes which are considered the main causes of change at the molecular level.

I.1.2 Sex chromosome systems

Bull (1983) reviewed the variety and distribution of sex chromosomes in nature. Generally, most sex chromosome systems are XX/XY; meaning that one sex is homogametic and carries two copies of the same sex

chromosomes (XX) and the other sex is heterogametic and carries just one copy of this chromosome, paired with another chromosome that differs in some way (XY). The difference could be as simple as the presence of a sex factor on Y and missing from X (or the absence on Y of a factor present on X) e.g. the Emu (Ogawa, Murata, and Mizuno 1998).

There is variation in the precise control of sex determination. In mammals, the presence of the Y chromosome (or more specifically a single sex determining gene - SRY in humans) dictates that the individual should develop as a male (Bull 1983; Berta et al. 1990). Mammalian aneuploids (atypical karyotype) featuring one Y chromosome but two or more X chromosomes (e.g. XXY, XXXY) still develop as males (Bull 1983). This is a 'dominant Y' system. However, in Drosophila and other dipterans, it is the ratio of X chromosomes to autosomes (A) that determines sex (Bull 1983). Females are produced when there are two X chromosomes for each normal set of autosomes and males when there is just one. This means that in this 'recessive X' system, the Y chromosome's only function in sex determination is to segregate opposite the single X during male meiosis (spermatogenesis). Accordingly, Drosophila aneuploids with a single X chromosome but no Y chromosome (XO) still develop as males (Bull 1983). XXY individuals develop as females because the number of Xs, and hence the X/A ratio, is the same as in normal females.

For most taxa where sex chromosomes have been found, it is as yet unknown whether the heterogametic system is dominant Y, recessive X or

some other mechanism. Where this has been done, dominant Y is the more common of the two mechanisms (Bull 1983) and it is believed that the recessive X system, or the X/Autosome balance, may evolve from dominant Y, following the full degeneration of the Y chromosome, a process I shall discuss in greater detail later on (Charlesworth 1996).

In both the mammals and flies, the heterogametic sex is the male. The same is true throughout most other insect orders, the arachnids, and, where there are separate sexes, the nematodes (Bull 1983). However, this is not a universal rule. Throughout the lepidoptera (butterflies and moths), the female carries the non-identical set of sex chromosomes and, closer to home, female birds do the same (Bull 1983). With female heterogamety the convention is to denote this system as ZZ for males and ZW for females to distinguish this female heterogamety from male heterogamety. Generally, the biological implications are presumed to be the same whichever sex is heterogametic and in much of the following I will only refer to the ZZ/ZW system when it has been shown to differ from the pattern observed in the XX/XY system. Although each system is conserved within several large and specious taxonomic groups there are as many examples of variation within groups at order, family, or even genus level.

For example, the sex chromosomes of mammals and birds evolved independently sometime after their divergence 300-350 MYA. In fact, the chicken Z sex chromosome, or at least a part of it, shows significant synteny in gene homology and order with human chromosome 9 (Nanda

et al. 1999) (See also Ellegren 2000). That the mammalian sex determining gene SRY is not sex-specific in birds or reptiles suggest that it gained this function after the mammalian divergence.

In amphibians, where genetic sex determination seems to be ubiquitous and sex chromosomes fairly widespread, some groups are still susceptible to environmental sex determination by temperature (Wallace, Badawy, and Wallace 1999). Interestingly there is good evidence that the amphibian ancestral ZZ/ZW state has been converted to an XX/XY state in at least 7 distinct lineages (Hillis and Green 1990). The existence of OW/OO females in the New Zealand frog species, *Leiopelma hochstetteri* suggests that the amphibian W chromosome determines sex by the presence of either an ovary determining or a testis-inhibiting factor. In the case of the latter, Hillis and Green (1990) postulate that the transition from ZZ/ZW to XX/XY could be facilitated by a recessive mutation of autosomal testis determining genes which would subsequently take over sex determination and become the sex determining locus (i.e. a new Y chromosome).

No clear distribution or trend seems to exist within the fish. At the order level, Perciformes contains examples of hermaphroditism and of sex chromosomal, temperature dependent, and multi-factorial sex determination (Chen and Yeung 1983). Both XX/XY and ZZ/ZW forms have been discovered in the same genus (e.g. Southern platyfish, *Xiphophorus maculates*, (Chen and Yeung 1983)) and in many groups where a single sex locus determines sex, the 'sex chromosomes' are often

undifferentiated (e.g. *Danio rerio* (Singer et al. 2002)). In medaka, *Oryzias latipes*, the male conferring effect of the small non-recombining region of the Y chromosome, can be substituted by autosomal polygenes in XX individuals to turn them male (Nanda et al. 2003). Surprisingly, sex determination in the well-studied 3-spine stickleback, *Gasterosteus aculeatus,* has only recently been found to be XY chromosomal (Peichel et al. 2004).

I.1.3 The transition from autosomes to sex chromosomes

For mammals and birds, it is very unlikely that chromosomal sex determination arose at the same time as separate sexes (i.e. from an hermaphrodite ancestor, see below). Ancestral vertebrate groups had separate sexes without sex chromosomes long before vertebrates came onto land (Bull 1983). Instead, it is likely that sex determination came under a single locus (analogous to SRY but not necessarily homologous to it). Once this had been established, this allele began to segregate like any other except that, by definition, it was never present in females. It may well be that this is as far as sex determination has gone in many groups; not yet sex chromosomal, just a single sex locus.

As far as sex determination goes, this is all that is required. However, in many organisms, something happened next. Recombination was in some way restricted between the chromosome carrying the sex factor and it's homologue. Chromosomal inversions are one way in which this could be

achieved. If the region around the sex factor was inverted, this would create a region on the developing Y that could no longer pair homologously with the X chromosome. Lahn and Page (1999) showed that differentiation of X and Y chromosomes in the mammalian lineage leading to humans featured four major inversions. These are now detectable as four 'strata' of X-linked genes showing internally similar degrees of divergence from their remaining Y homologues (see Figure I.1). Three of these strata are shared by mouse X chromosomes (Sandstedt and Tucker 2004), though the fourth seems to have originated independently in different mammalian lineages (Iwase et al. 2003). In the ancestor of all Old and New World monkeys and apes, the inversion moved the sex determining locus SRY into the previous pseudo-autosomal region (PAR) (Glaser et al. 1999), the proximal portion of which now lay in the non-recombining region (NRY). This means that recombination was 'turned off' in four stages preceding its almost complete loss from the human Y chromosome. Furthermore, the development of dosage compensation of the X (see below) most probably progressed by stages with groups of X-linked genes becoming susceptible to inactivation at similar times. The existing PAR now accounts for just a small percentage of the sex chromosomes. As recombination must occur here at least once in every meiosis (Rouyer et al. 1986), the recombination rate is extremely high. The evolution of the genes in this region is the subject of Chapter II. At the other extreme, some genes have survived on X and Y despite tens of millions of years without recombination, these are the subject of Chapter III. On the border between the older and highly degenerate strata (1, 2 & 3), and the actively degenerating stratum 4, sits AMELX, which, along

with its Y homologue AMELY, is the subject of Chapter IV. As the strata model is becoming accepted for mammals, it seems also to explain the cessation of recombination in both birds (Ellegren and Carmichael 2001) and plants (Filatov, submitted). A gene from the sex chromosomes of plants in the genus Silene is the subject of Chapter V and their sex chromosomes are described there.

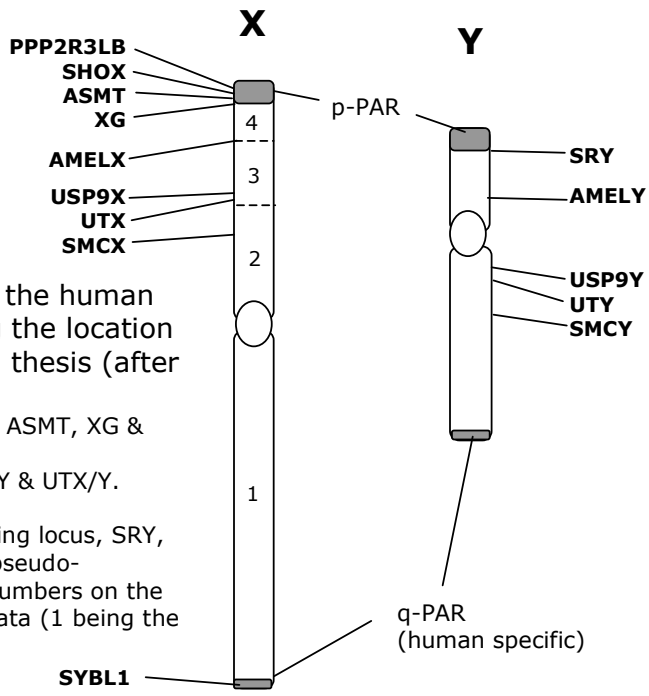


Figure I.1 – Schematic of the human sex chromosomes detailing the location of the genes studied in this thesis (after Lahn & Page 1999).
Chapter II: PPP2R3LB, SHOX, ASMT, XG & SYBL1.
Chapter III: SMCX/Y, USP9X/Y & UTX/Y.
Chapter IV: AMELX/Y
The position of the sex-determining locus, SRY, is also given. Shaded areas are pseudo-autosomal and can recombine. Numbers on the X chromosome correspond to strata (1 being the oldest).

As an alternative to chromosomal inversion, a gene controlling recombination could mutate and restrict recombination in its vicinity or, as in Drosophila, 'turn-off' all recombination in the heterogametic sex (Drosophila males). Recently, Kondo *et al.* (2001) showed that for the largely homologous and undifferentiated sex chromosomes of the fish Medaka, *Oryzias latipes*, recombination was only restricted in the

immediate vicinity of the sex determining locus. Whichever way recombination is restricted, there really has to be a third gene involved – something to encourage sex-specificity of an entire chromosome (or a good part of it). So, the three requirements for the differentiation of the sex chromosomes are (i) a sex-determining locus, (ii) a linked locus restricting recombination and (iii) a linked locus with a *cis* allele advantageous sex-specifically in the heterogametic sex. If a region with reduced recombination had already been created, the third requirement may more easily be added later by the translocation of a sex-specific gene from an autosome.

The most plausible and best supported, mechanism yet proposed for the creation of a sex chromosome system came from Charlesworth and Charlesworth (1978). Focusing on flowering plants, their model progresses from a co-sexual, hermaphroditic population, through gynodioecy (females and hermaphrodites) and on to dioecy. Most probably this would begin with a recessive mutation for male sterility. Recessive because that is the most common type of new mutation and for male sterility because, in competition with hermaphrodites, there is more reproductive success in having ovules fertilised by other individuals than in trying to compete in the pollen war. Once a male-sterility mutation has invaded the population, its frequency will be limited to less than the frequency of remaining hermaphrodites because they are the only source of pollen. However, in this situation there may now be an increased advantage for a female-sterility mutant (i.e. males) if individuals with this mutation can produce significantly more pollen than hermaphrodites. Having both sterility

factors would cause full sterility, so there will be strong selection for linkage to separate the two mutants. Nearby loci on homologous chromosomes would be ideal. In this mechanism, crossover suppression and the formation of sex chromosomes is closely linked to the creation of separate sexes (dioecy).

This system is the subject of Chapter V and, though in general agreement with observations from animal sex chromosomes, the genetic content of Silene sex chromosomes will be described there. In the genus Silene, most of the around seven hundred species are hermaphroditic or gynodioecious (Desfeux et al. 1996). However, six species have evolved with dioecy and chromosomal sex determination has been detected amongst these. (Desfeux et al. 1996; Gutman and Charlesworth 1998). The number of hermaphroditic and gynodioecious species in this genus strongly promotes the above mechanism. What is especially interesting is that this change happened recently, probably within the last 20 MY. Genes on this Y chromosome also seem to have a higher mutation rate than the X (Filatov and Charlesworth 2002) and yet show 20-fold lower diversity (Filatov et al. 2000). One functional X linked gene has been found to have a Y gametologue with an early frameshift indel, which probably inactivates this gene. Importantly though, this is still a young pair of sex chromosomes compared with those of mammals, birds or Drosophila. Though the Y chromosome has altered morphologically, becoming physically larger than the X, it is still largely euchromatic (Siroky, Castiglione, and Vyskot 1998). Studying the sequence and genes of these

sex chromosomes could give us a very good idea of the predominant forces shaping the early evolution of sex chromosomes.

I.1.4 Restriction of recombination

The restriction of recombination is the key process in the evolution of sex chromosomes. Without it, sex factors may continue to segregate at a locus but sex chromosomes will not differentiate, as can be seen in the ancient but still homomorphic sex chromosomes of the emu (Pigozzi and Solari 1997; Ogawa, Murata, and Mizuno 1998). As soon as recombination has been restricted in some part of the genome, a novel situation arises for traits that would increase fitness of the heterogametic sex but confer a disadvantage if expressed in the homogametic sex (Rice 1992; Rice 1996). As pointed out by Bull (1983), at this point the evolution of sex chromosomes detaches from that of sex determination and such chromosomes begin to gain special characteristics.

Some genes benefit males but confer a disadvantage if expressed in females, especially in relation to sexual selection – bright colours, large ornaments, heavy weapons (Rice 1992). It is possible that many of these could be turned on in a sex specific manner whilst on autosomes. However, if one became linked to a sex factor on a fledgling sex chromosome, it would fill the third requirement for the promotion of recombination restriction. Location on the Y does seem to confer a benefit to the genes for male specific tissues. Indeed, a relative abundance of

these genes have been found on the Y chromosomes of mammals (Delbridge and Graves 1999) and Drosophila (Rice 1987; Carvalho et al. 2001), and appear to have been translocated from autosomes to the Y (Lahn and Page 1997; Carvalho 2002). Probably due to a lack of study, the same pattern has yet to be shown for a female specific W chromosome, though the first female specific gene from the chicken W, has recently been characterised (Ceplitis and Ellegren 2004). Interestingly, it has been hypothesised, that of the two main sex chromosomal systems, the ZZ/ZW system is more suited to the rapid evolution of sexually selected traits, and that this may explain the preponderance of 'showy males' in birds and butterflies (Reeve and Pfennig 2003).

Lahn and Page (1997) developed the idea of "functional coherence" of the mammalian Y chromosome. They managed to divide genes in the non-recombining region (NRY) into two distinct categories. One group have active X homologues, are expressed in many different tissues and are present as only one copy on the Y chromosome. In contrast, the second category are genes with no X homologue, which are specifically expressed in the testis and, interestingly, which have multiple copies on the Y chromosome. This last point particularly emphasises the likely sex-specific advantage of these genes. A more recent review by Lahn, Pearson and Jegalian (2001) reveals the situation to be a little less clear cut with a third class of Y-linked genes of intermediate or unknown allegiance to the two previously described groups. However, their basic proposal of functional coherence holds and is strongly supported by work on

Drosophila. Carvalho et al.(2000) (see also Carvalho et al. 2001) showed a very distinct clustering of genes on the fruit fly's Y chromosome that are involved in male fertility, specifically molecular motor proteins active in sperm. In addition, the absence of evidence of X-Y gametologous genes on the Drosophila sex chromosomes fits the hypothesis that the current Drosophila Y chromosome is not the original partner to the X (Carvalho 2002).

I.1.5 Dosage Compensation

For reasons that I will discuss below, Y chromosomes have few functional genes. The vast majority of genes on the human X chromosome are left without a functional homologue on the Y chromosome (a gametologue (Garcia-Moreno and Mindell 2000)). Current numbers of genes for the human X and Y chromosomes are 1141 and 255 respectively (Genbank Genome Build 35.1: Benson et al. 2002). Given that a diploid genome contains two functional copies of each autosome, the translational machinery of the cell has evolved to 'expect' bi-allelic expression. So, from the cell's perspective, many X-linked genes are under-expressed in males.

Again, the mammals and Drosophila can be used as examples of nature's two main counter measures against such an effect. In ourselves and our furry relatives, expression from X chromosomes is altered in females such that one entire X chromosome is inactivated. In placental mammals

(Eutheria), either X chromosome has an equal chance of inactivation, determined on a cell-by-cell basis. However, in the marsupials, the X inherited paternally is always singled out for inactivation (Bull 1983).

In mammals, the inactivation of the X chromosome spreads as a signal emanating from a single X-linked locus, XIST. This gene does not encode a protein, but a non-coding RNA (ncRNA) and this is antagonised by a reverse transcript of itself, *Tsix* (Ogawa and Lee 2002). Inactivated genes are condensed and become heterochromatic (facultative heterochromatin).

Genes in the PAR at the end of the short arms of the human sex chromosome (the p-PAR, Figure I.1) escape inactivation, probably because, being present on X and Y chromosomes, their dosage is already equal in males and females. The genes of the second pseudoautosomal region (q-PAR), which is specific to humans, however, are inactivated. This is probably carried over from their relatively recent translocation from the X chromosome (Ciccodicola et al. 2000). Interestingly, it seems some X-specific genes are still required in two copies. In females, there are around ten X-linked genes, which escape inactivation to be expressed from both chromosomes (Disteche 1995; Lahn and Page 1997; Ellis 1998). In males, the gametologues of these genes are still functional on the Y chromosome and form a distinct class of single-copy, widely expressed genes in contrast to the other class of multi-copy, male-specific Y-linked genes (Lahn and Page 1997). In '45, X' aneuploid individuals, it is thought that the lack of a second copy of these genes is responsible for

the phenotype of Turner's syndrome (Zinn, Page, and Fisher 1993). There is a strong correlation between the absence of a Y homologue and X inactivation – evidence that Y degeneration is the driving force for X inactivation (Jegalian and Page 1998).

In Drosophila, X-linked genes are expressed in females as per the autosomes, but expression is doubled on the single copy X chromosome of males (Charlesworth 1996). The few successful studies of this phenomenon in birds, as reviewed recently by Ellegren (2002), point to the presence of dosage compensation and in a form unlike that of mammals or Drosophila. Of the Z linked genes that have been studied, complete inactivation of one chromosome has not been found and preliminary transcriptional studies point to a down-regulation of Z activity from both chromosomes in males (Kuroda et al. 2001).

I.1.6 Y degeneration

Given the huge natural variety of modifications to the basic heterogametic sex chromosome system, it is all the more striking that there are strong similarities in the morphology, genetic structure and sequence composition between sex chromosomes from very distantly related taxa.

As I have previously alluded to but not explained, Y chromosomes have relatively few functional genes. The high number of pseudogenes with active X homologues found on the human Y chromosome and the

presence of dosage compensation support the idea that the Y has lost genes rather than the X gained them. In virtually all taxa studied, more X-linked genes have been found than Y-linked (Bull 1983). When stained, Y chromosomes often appear heterochromatic, the usual symptom of gene poor regions. Additionally, Y chromosomes accumulate both repetitive DNA sequences (e.g. Silene, though the Y is still euchromatic - Siroky, Castiglione, and Vyskot 1998) and non-functional retroviruses (Kjellman, Sjogren, and Widegren 1995). Such sequences could increase the likelihood of chromosomal inversions via illegitimate recombination. In taxa where it has been possible to force sex-reversed genotypes (e.g.females with genotype XY), YY and Y0 are often inviable, suggesting an absence of vital genes (Bull 1983).

The speed of the degeneration process, for Drosophila species at least, is slow in population genetics terms. The fusion of sections of autosomes to ancient sex chromosomes (forming neo-X or neo-Y chromosomes) in several Drosophila groups show different stages of degeneration as the previously ancestral sequence succumbs to the effects of life on the Y (Charlesworth 1996). A fusion approaching 13 million years of age in the *D. pseudoobscura* species group shows complete inactivation whereas a younger fusion in *D. miranda*, probably less than 2 million years old shows substantial but not complete inactivation. A very young fusion that occurred in the last few thousand years in *D. americana* retains the majority of its functionality (Charlesworth 1996).

These properties of Y are found in each independent evolution of sex chromosomes in mammals, birds (W), Drosophila, Silene and pretty much any group with chromosomal sex determination. Together they are known as Y degeneration and this process is the central concept in the molecular evolution of sex chromosomes. The most interesting aspect of X chromosome biology, dosage compensation, is a reaction to degenerative changes on the Y chromosome (Ellis 1998). The convergent evolution of Y degeneration in different taxa strongly suggests that a single mechanism operates. Despite considerable interest and theorising, it is still unclear which, if any, of the forces discussed in the next section predominate (Charlesworth and Charlesworth 2000).

I.2 The forces guiding the molecular evolution of sex chromosomes

I.2.1 Genetic processes in regions of low recombination

Whilst it is clear that recombination increases the efficacy of selection (Felsenstein 1974; Kliman and Hey 1993; Betancourt and Presgraves 2002) and in its absence, selection may be too inefficient to remove deleterious mutations, precisely how this happens and under what circumstances, is unknown. The following mechanisms are all thought to contribute in part to the degeneration of Y chromosomes due to the lack of recombination.

I.2.1.1 Muller's Ratchet

Muller (1932) proposed a mechanism whereby non-recombining chromosomes (e.g. the Y chromosome) show a progressive increase in the average number of slightly deleterious mutations per individual due to the stochastic loss of the class of chromosomes with the least number of deleterious mutations. Considering all the mildly deleterious mutations in the population, some chromosomes will have more than others. Chromosomes with very many deleterious mutations will be removed by selection, but these will be replaced with new mutations to chromosomes carrying slightly fewer mutations. At the other end of the spectrum, some chromosomes will contain very few (even zero) deleterious mutations.

However, if the selective effect of deleterious mutations is low, then the healthiest class of chromosomes will only form a fraction of the population. By chance, this group may be lost from the population in a single generation. If this happens in a recombining population, the healthiest class may be reconstituted from healthy parts of surviving chromosomes. However, in a non-recombining population, the healthiest class of chromosomes is lost forever (assuming the rate of back mutation is low).

Because of the irreversible nature of this process, it was named Muller's Ratchet by Felsenstein (1974). Its effectiveness depends on the mean number of deleterious mutations on Y chromosomes, the strength of selection against mutations and on the effective Y chromosome population size. For Drosophila, where the fittest class of Y chromosomes is likely to number in the thousands, the effectiveness, or 'speed', of Muller's Ratchet is likely to be low (Charlesworth and Charlesworth 2000; Gordo and Charlesworth 2000; Gordo and Charlesworth 2001).

Since Muller, a range of related mechanisms involving selection, rather than drift, has been proposed. Reviewed by Charlesworth and Charlesworth (2000), genes on a non-recombining Y chromosome, may be subject to the following evolutionary forces.

I.2.1.2 Background Selection

The fixation of neutral, or slightly deleterious mutations depends largely on the behaviour of the Y chromosome as a whole unit. Without recombination, the removal of a deleterious mutation by selection from the population of Y chromosomes, will also remove all other diversity present on the same chromosome (Charlesworth, Morgan, and Charlesworth 1993). With recurrent deleterious mutations, the reduction in diversity may severely curb the adaptive potential of the Y chromosome population. Interestingly, the reduction in diversity estimated under background selection may resemble a reduction in effective population size (Charlesworth, Charlesworth, and Morgan 1995). In turn, this may promote the action of Muller's Ratchet.

I.2.1.3 Genetic hitch-hiking

By linkage with a strongly selected beneficial mutation, slightly deleterious mutations may 'hitch-hike' to fixation as the beneficial mutation supplants all competitors in a selective sweep (Smith and Haigh 1974). In the extreme, all population diversity will be over-written. The accumulation of male-specific genes on the Y chromosome, especially those involved in reproduction (Lahn and Page 1997), may be a strong contributing factor to its degeneration. Reproductive genes evolve extremely fast (Wyckoff, Wang, and Wu 2000; Van Doorn, Luttikhuizen, and Weissing 2001; Swanson and Vacquier 2002; Swanson, Nielsen, and Yang 2003), and as these genes sweep to fixation, other loci, perhaps partially sheltered by the X linked copy, may begin to accumulate deleterious mutations.

I.2.1.4 Hill-Robertson weak selection interference

As proposed originally by Hill and Robertson (1966), weakly selected alleles on different copies of the same chromosome, will be seen together by selection. If two beneficial mutations occur at the same locus, in the same population, without recombination only one may go to fixation. What is more, if the difference between them is small, the time taken for the winner to go to fixation will be greater than if it had only had to compete against a single ancestral allele (McVean and Charlesworth 2000). This argument can be extended to mutations occurring at two different linked loci, as long as the rate of recombination between them is low. If both beneficial alleles do not recombine into a single *cis* haplotype, then they will compete. Finally, and of particular applicability to Y chromosomes, a new weakly selected deleterious mutation occurring in cis (associated) with a weakly beneficial mutation, will cause a similar effect. This is because the ancestral allele at the second locus (featuring the deleterious mutation) will now be a weakly beneficial polymorphism in opposition to the weakly beneficial mutation at the first locus.

In the Drosophila genome, Hey and Kliman (2002) estimated that this effect would not be significant in regions with a local recombination rate greater than 1.5cM/Mb. On the Y chromosome, this is taken to the extreme with a large number of loci unable to recombine. This process lessens the efficacy of selection to act against slightly deleterious mutations, allowing them to accumulate in non-recombining regions.

Evidence for this process has come from Drosophila, in which codon bias is reduced in areas where Hill Robertson Interference would be more effective (Kliman and Hey 1993; Hey and Kliman 2002) and the length of introns is increased in regions of low recombination, an adaptation which would increase the local recombination rate per locus (Comeron and Kreitman 2002).

I.2.2 Which process is predominant?

All the above processes are difficult to distinguish; they are not mutually exclusive and give similar predictions over changes in diversity and the efficacy of selection. Some differences have been proposed, mainly in the theoretical literature.

Positive selection and background selection both predict the observed reduction of diversity in regions of low recombination but in addition, positive selection will skew the frequency spectrum of polymorphisms such that most polymorphisms are singletons (i.e. recently derived since the selective sweep) (Andolfatto 2001). For the neo-Y chromosomes of *Drosophila miranda* (formed by a translocation of a section of autosome to the original Y chromosome) a signal better matching the hitch-hiking hypothesis has indeed been observed (Bachtrog 2004). However, as the author notes, the sign of any other process acting in the past, would have been over-written by the selective sweep.

The stochastic nature of Muller's Ratchet could aid in distinguishing this process from others driven by selection. From Charlesworth (1996) :

> "…if the background selection or selective sweep models apply, one would expect to see an accelerated rate of amino acid replacement substitutions, and of silent-site substitutions that change preferred to non-preferred codons. This is not a requirement of Muller's Ratchet."

In addition, each of the processes depends on the deleterious mutation rate so may be expected to occur faster in younger chromosomes than older ones or in large non-recombining regions rather than smaller. Hence it is also important to know how the mutation rate may vary.

I.2.3 Variations in mutation, diversity and substitution

The search for a clear explanation of how natural selection, mutation and drift shape molecular evolution has been considerably complicated by efforts to measure the variation in mutation rate between the sex chromosomes and across the genome.

One of the first potential sources of variation identified was a male derived bias in the rate of novel mutations, which would in turn bias substitution rates (under neutrality) and diversity, and explain the occurrence of several diseases along the way. Crow (2000b) cites Weinberg (1912) as the discoverer of the increased likelihood of some genetic diseases with

paternal age effect, and Haldane (1947) with the discovery that mutations causing deleterious haemophilia phenotypes, arose more often in the male germline. Haldane conceived 'male driven evolution', the effect whereby the male germ line, having more mitotic replications (a known source of mutation), contributes more mutations, and therefore more variation on which selection can act (Crow 2000b). The subsequent elucidation of the genetic code and the development of the molecular theory of evolution, in particular, Kimura's neutral theory (1983), revealed the potential to measure mutations quantitatively at the level of DNA. Miyata *et al.* (1987) proposed a model for male-driven molecular evolution. Importantly, for the genes on the sex chromosomes, this bias would give genes on the Y chromosome a potentially greater rate of evolution.

Though intuitively the male-mutation bias hypothesis makes sense, and disease linkage lends strong medical evidence for it (Crow 2000b), the molecular evidence has been conflicting. In complex eukaryotes, it is difficult to measure the mutation rate directly and most studies have used the substitution rate between species or the level of diversity within a species. Estimates of the ratio of male to female derived mutations (α) have varied depending largely on the methods used or the genes studied (see review by Hurst and Ellegren 1998; Bohossian, Skaletsky, and Page 2000; but see also - Crow 2000a). Interestingly, very different estimates of α were obtained in each of the comparisons between X and autosomes, between Y and autosomes and between X and Y chromosomal genes. Table I.1 lists some of the estimates that have been obtained from different sequence samples in a (mammalian-biased) sample of studies.

Table I.1 – Estimates of the male mutation bias from various studies

Taxa	Data	α	Study
Humans	A 39kb X to Y transposition	1.66 (95% C.I. 1.15-2.87)	(Bohossian, Skaletsky, and Page 2000)
Apes	18 pseudogenes	3.6 (range 1- ∞)	(Nachman and Crowell 2000)
Primates	AMELX/Y, SMCX/Y & ZFX/Y genes	5.06 (95% C.I. 2.42 - 16.6)	(Huang et al. 1997)
Primates	ZFX/Y genes	~6	(Shimmin, Chang, and Li 1993)
Apes	10kb homology between Chr 3 and Y	5.25 (95% C.I. 2.44- ∞)	(Makova and Li 2002)
Goats & Sheep	ZFX and ZFY	2.93 (95% C.I. 1.51-8.61)	(Lawson and Hewitt 2002)
Cats	ZFX and ZFY	4.5	(Pecon Slattery and O'Brien 1998)
Birds	ATP5A1 gene on X, Y and autosome	1.8-6.5 in different bird lineages	(Carmichael et al. 2000)
Salmonid fish	GH- 2 gene from Y and autosome	5.35-6.6	(Ellegren and Fridolfsson 2003)

No male mutation bias has been detected in Drosophila, probably because there is no great difference in the number of cell divisions required to create sperm or eggs (Bauer and Aquadro 1997). Whilst, newer estimates of α from more taxa continue to show a consistent male bias (Table I.1), it appears that the mutation rate also varies at other levels within the genome (Smith and Hurst 1999).

Originally, Mouchiroud, Galtier and Bernardi (1995) found that individual genes had specific mutation rates based on silent divergence (which should be neutral and proportional to mutation rates). Matassi et al. (1999) expanded this to pairs of neighbouring genes, which had more

similar rates of synonymous substitution than expected by chance. Pseudogenes, which should be free from selection, when translocated into a new part of the genome, begin to mutate to resemble their surroundings in base composition (Casane et al. 1997). Lercher, Williams and Hurst (2001) found that the correlation extended over a broader range, perhaps even that the synonymous substitution rate, was governed, at least in part at the chromosomal level. Still, the silent substitution rate in adjacent blocks of sequence is correlated within human chromosome 7 (Smith, Webster, and Ellegren 2002). McVean and Hurst (1997) found that X chromosomes may experience a reduced rate of substitution relative to other chromosomes in response to hemizygous expression in males and that this was not due to a relative increase in mutation rate on the Y (due to male mutation bias). However, Malcom, Wyckoff and Lahn (2003), using substantially greater sequence, found that, whilst the X chromosome was an outlier, it did not fall outside the range predicted from a male mutation bias for a chromosomes spending a disproportionate amount of time in females. This latter study also found that regions of similar substitution rate (synteny blocks) were clustered on chromosomes.

Silva and Kondrashov (2002) found the blocks of similarity to be in the range of 1000-10000bp using human-baboon divergence data. Using diversity data, Reich et al. (Reich et al. 2002) proposed that the observed regional variations are better explained by shared genealogical history than blocks of similar mutation rate. Their estimate of the range of this variation was tens of thousands of bases. They proposed that large

regions containing substantial linkage disequilibrium are separated by hotspots of recombination, which reduce disequilibrium between adjacent regions. This has been supported by a recent recombinational map showing that 80% of recombination events occur within 25% of the sequence (McVean et al. 2004).

Whilst there seems to be heterogeneity between chromosomes, the distance over which this variation acts has yet to be fully determined and the heterogeneity amongst different chromosomes, which increases with decreasing chromosome size (for the autosomes), may be a natural sampling effect.

I.2.4 Correlates of recombination

Whilst diversity was used in some of the above studies to measure the local mutation rate, it is rather susceptible to various population genetic processes, including those acting in regions of low recombination. With the recent increase in availability of human sequence and better estimates of genomic variation in recombination rate, it is becoming clear that diversity too may vary as discrete regions within the genome. Again, the relative importance of selection, mutation and drift have yet to be disentangled.

Nachman et al. (1998) found an order of magnitude difference in diversity between different loci along the X chromosome. However, diversity is

generally low in humans (Kaessmann et al. 2001) and the range of X linked diversity estimated was 0-0.2%. As diversity cannot drop below zero, there may be a strong edge effect, which causes small increases above this to look like big differences.

Diversity on Y chromosomes is consistently low across different taxa (Filatov et al. 2000; Shen et al. 2000; Hellborg and Ellegren 2004) and yet divergence is high (Wyckoff, Li, and Wu 2002). Whilst in Drosophila, the diversity spectrum may signify the presence of selective sweeps acting on the neo-Y chromosomes (Bachtrog 2004), in humans, recent population expansion seems to have been a strong determiner of the lack of diversity (Kaessmann et al. 2001).

Across the genome, clustering of highly conserved, house-keeping genes may explain some of the variation in recombination rate. Tight clusters will be more affected by Hill-Robertson interference, but if they are essential genes, then there will be relatively few deleterious mutations of small effect (Pal and Hurst 2003). Selection can encourage recombination in regions where novel mutations are more likely and deleterious mutations may otherwise drift in the population for some time.

Lercher and Hurst (2002) measured a positive correlation between recombination and single nucleotide polymorphism (SNP) diversity across the genome in areas quite distant from known coding regions. They reasoned that any effect of positive or negative selection should be

negligible and that a systematic mutation bias was operating, perhaps caused by recombination itself. This is the subject of Chapter II.

II HIGH SUBSTITUTION RATES IN MAMMALIAN PSEUDOAUTOSOMAL GENES

Dave T. Gerrard

School of Biosciences,

The University of Birmingham,

Edgbaston, Birmingham, B15 2TT, UK

II.1 Introduction

Genetic recombination increases the efficacy of natural selection, pushing the advantageous mutation toward fixation and helping weed out the deleterious (Felsenstein 1974 and see Chapter I). But could recombination also be a source of mutations? Strathern et al. (1995) found a hundred-fold increase in the mutation rate of a yeast strain in which a high rate of recombination was induced. The greatest density of recombination events yet observed for the human genome was in the pseudoautosomal region (PAR) of the sex chromosomes over a stretch of 2.6 Megabases (Mb) (Lien et al. 2000). Lacking significant homology along much of their length, the X and Y chromosomes are forced to pair in this region in order to segregate properly. This means that at least one chiasma must form in this relatively short stretch of DNA at every male meiosis (Burgoyne 1982; Hale 1994). Using high resolution sperm typing in this region Lien et al. (2000) measured a recombination rate greater than 20 times the genomic average of ~1cM/Mb, far above even the supposed 'jungles' of high recombination elsewhere in the genome (Yu et al. 2001). If recombination is mutagenic, then this is surely the place to look. Yet, without the means to directly measure the mutation rate in a specific region of the human genome, we need some other signal that would be correlated with the amount of mutations. The Neutral Theory (Kimura 1983) predicts that in the absence of selection, the substitution rate will equal the mutation rate. With increased mutagenesis, we should see significantly higher divergence of the neutral sites in the PAR compared to the rest of the genome.

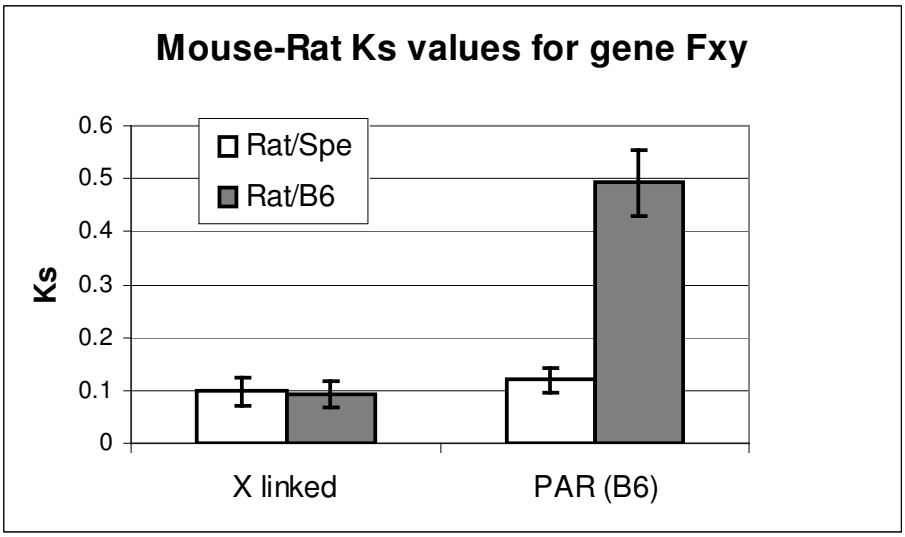


Figure II.1 - Effect of location in PAR on synonymous site divergence for gene Fxy from rodents. Rat/Spe are divergence values for the Rat - *Mus spretus* comparison. Rat/B6 are the data for the Rat - *Mus domesticus* comparison. X-linked refers to the section of Fxy which is X linked in all 3 species. PAR(B6) refers to the section which is pseudo-autosomal in *Mus domesticus* only.

Support for this idea has already come from a gene spanning the boundary between the PAR and the differential sex chromosome regions in the laboratory mouse, *Mus domesticus* (Perry and Ashworth 1999). Conveniently, the same gene is fully X–linked in another mouse, *Mus spretus*, allowing Perry and Ashworth to make a within-genus comparison of the effect of elevated recombination on a gene. When the region that is X-linked in both species was compared with rat, the two mouse species showed an almost identical number of silent site differences. However, the region that was pseudoautosomal in *M. domesticus*, showed much greater divergence from rat than the homologous, yet X-linked, region in *M. spretus* (See Figure II.1).

We wanted to investigate how widespread this effect is. We compared the silent divergence values for neutral sites in the PAR of closely related

species and a distribution of similar values genome wide. The latter somewhat limited the range of investigation to mouse-rat comparisons and human with another primate because of the quantity of sequence available in public databases. We used orangutan because it represented the best balance of divergence and availability of sequence. Accordingly, I assembled a distribution of human-orangutan divergence values from available orthologous sequence data and used this to show that the divergence of genes in the PAR (newly sequenced by D. Filatov) was exceptionally high, suggesting a greater mutation rate operating in this region.

II.2 Materials and Methods

II.2.1 DNA sequencing

D. Filatov provided the sequences listed in Table II.1. The genes are listed in order of distance from the X/Y p-arm telomere. The gene XG spans the PAR boundary with its more distal (from the centromere) sequence in the PAR and its proximal sequence specific to the X. The SYBL gene lies at the distal end of the X chromosome q arm which is specific to the X in apes (i.e. the orangutan) but is pseudo-autosomal in humans as we possess a second PAR (q-PAR) that originated since the divergence of men from apes (Ciccodicola et al. 2000). For the purposes of this study, coding sequence was removed from any analyses.

Table II.1 – Location of sequences obtained from orangutan for this study. Physical locations from following sources: PPP2R3LB, (Schiebel et al. 2000), 2000; SHOX, (Rao et al. 1997); ASMT, (Rodriguez et al. 1994); XG, (Rappold 1993); SYBL1, NCBI map of the human X chromosome, build 29. Recombination data from (Lien et al. 2000) except SHOX(+15/-17), (May et al. 2002) and SYBL1, (Ciccodicola et al. 2000).

Gene (or region thereof)	Location	Distance from p-telomere	Recombination Rate (cM/Mb)
PPP2R3LB	p-PAR	150kb	15
SHOX (+1/-12)	p-PAR	500kb	27
SHOX (+15/-17)	p-PAR	500kb	200-350
SHOX (+7/-9)	p-PAR	500kb	27
ASMT	p-PAR	1000kb	36
XG (+1/-2)	p-PAR	2.6Mb	26
XG (PABxy)	p-PAR	2.6Mb	26
XG (PABx)	X-linked	2.6Mb	0
SYBL1	q-PAR	151.7Mb	6

II.2.2 Building a distribution of human-orangutan divergence

Whilst estimates of human-orangutan divergence have been published (Chen and Li 2001), the distribution of such values from different regions of the genome has not. To compare the PAR genes with the rest of the genome, such a distribution was assembled. A list of all GenBank (Benson et al. 2002) entries sourced from the genus Pongo (both Orangutan sub-species) was obtained using Entrez at the NCBI web-site. All-against-all comparisons were then made using FASTA (Wisconsin Package, GCG) and all duplicate sequences (100% identity) were removed to leave 450 different sequences (though in many cases the differences were very slight). Each of these sequences was then BLAST-searched against GenBank. Where precisely the same piece of human sequence was matched equally well by two or more orangutan sequences, the longest orangutan sequence was kept. 204 orangutan sequences with putative human homologues remained. As the previous BLAST search could have left us with very similar Orangutan sequences matching the same human sequence but under a different accession, we then BLAST-searched these sequences against RefSeq[1], the non-redundant set of human genome sequences. This process automatically removed sequences containing human repeats (45 sequences). After again filtering for orangutan sequences sharing the same best match, we were left with 91 good homologous pairs. Of these, 51 sequences had non-coding regions over 100 nucleotides in length and are shown in Table II.2. Sequences were

[1] http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html

aligned using ProSeq 2.9 (Filatov 2002) and checked by eye. Divergences were calculated as the proportion of differences.

II.2.3 Mouse-rat PAR divergence

Similarly, we also looked at the PAR divergence between mouse and rat using the only sequenced rodent gene fully in the PAR, *STS*. Non-coding and non-genic mouse and rat sequences are more diverged than human/orangutan sequences so we used the rate of synonymous change (K_s) at coding sites as a measure of neutral evolution. The data of Wolfe and Sharp (1993) were used to represent the distribution of K_s values across the genome. Mouse and rat translated STS coding regions were aligned using CLUSTAL W (Thompson, Higgins, and Gibson 1994) and this protein alignment was used to align the nucleotide sequence by eye in ProSeq. The K_s value for STS was also calculated in ProSeq using the same method as Wolfe and Sharp (1993), that of Li, Wu and Luo (1985).

Table II.2 – Sequences aligned to build distribution of human-orangutan divergences.

Orangutan accession	Human RefSeq Contig	Human Chromosome	Length of aligned non-coding DNA	Substitutions in alignment	% Divergence
m69167	NT_026943	1	212	2	0.94%
ab037486	NT_019273	1	440	6	1.36%
af310681	NT_029874	1	9025	212	2.35%
ab003314	NT_021877	1	509	12	2.36%
m69172	NT_026943	1	163	5	3.07%
af229806	NT_005375	2	179	2	1.12%
x91111	NT_005403	2	1120	15	1.34%
af378147	NT_005265	2	1238	44	3.55%
u77650	NT_005079	2	387	17	4.39%
ab041383	NT_005927	3	113	2	1.77%
af252561	NT_005825	3	927	20	2.16%
u42764	NT_007592	6	576	11	1.91%
af179702	NT_007592	6	228	5	2.19%
af179694	NT_007592	6	762	24	3.15%
af179710	NT_007592	6	1049	44	4.19%
ab041365	NT_007819	7	333	6	1.80%
ab037500	NT_033968	7	718	18	2.51%
af003996	NT_008646	10	471	6	1.27%
af149711	NT_009325	11	332	1	0.30%
af005651	NT_009151	11	300	2	0.67%
af205416	NT_009325	11	1349	31	2.30%
af005647	NT_009151	11	383	10	2.61%
x05035	NT_009325	11	4514	122	2.70%
m18796	NT_009325	11	3155	89	2.82%
m21825	NT_009325	11	1701	48	2.82%
m18038	NT_009325	11	6918	196	2.83%
af333022	NT_009151	11	299	9	3.01%
af205408	NT_009325	11	1256	45	3.58%
af003974	NT_009743	12	372	5	1.34%
af003973	NT_009743	12	290	4	1.38%
af003975	NT_009743	12	422	9	2.13%
af375643	NT_024413	12	2871	81	2.82%
af092833	NT_009731	12	751	29	3.86%
m30682	NT_010194	15	565	15	2.65%
af084197	NT_010362	15	5083	207	4.07%
af215710	NT_010419	16	500	12	2.40%
af215714	NT_010419	16	632	20	3.16%
af116784	NT_010783	17	1023	30	2.93%
af127780	NT_011296	19	818	38	4.65%
ab041417	NT_011512	21	672	15	2.23%
af291652	NT_011520	22	9777	301	3.08%
af085462	NT_027217	X	771	11	1.43%
af280911	NT_011726	X	845	13	1.54%
u88979	NT_011812	X	1409	28	1.99%
af280903	NT_011588	X	839	21	2.50%
af279924	NT_011757	X	2976	79	2.65%
af280893	NT_025965	X	906	25	2.76%
ab041369	NT_011611	X	310	9	2.90%
af280922	NT_011786	X	871	26	2.99%
af279925	NT_011757	X	1385	46	3.32%
af279926	NT_011757	X	987	35	3.55%

II.3 Results

A total of 74130 non-coding sites from 51 different orangutan sequences were aligned with their human homologues. There were 2116 differences giving an average divergence of 2.85% (Standard deviation 0.061%). The non-coding regions of the three most distal of the newly sequenced orangutan genes showed significantly high divergence relative to the other sequences compared (See Fig. II.2, Table II.3). Similarly, the rodent PAR gene, STS, is an outlier to the distribution of 319 K_s values obtained by Wolfe and Sharp (1993) with a high K_s of 0.814 (±0.07) using the same method (Li, Wu, and Luo 1985) (See Fig. II.3). Other methods give slightly lower K_s values for STS but none approach the main distribution of K_s values (Modified Nei Gojobori (Jukes Cantor) 0.642 (±0.051); Pamilo

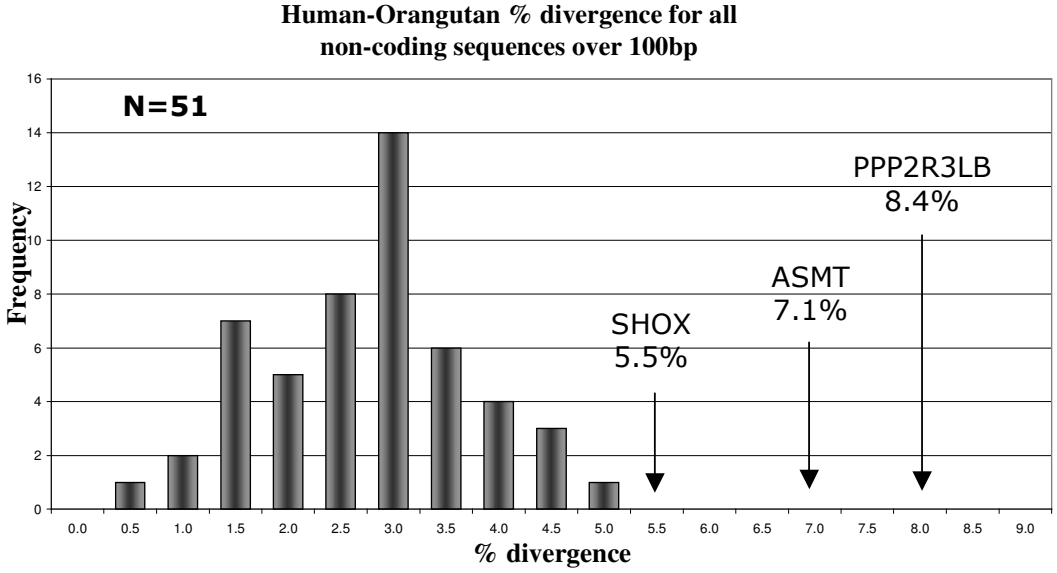


Figure II.2 Frequency distribution of non-coding sequence divergence from 51 human-orangutan homologous pairs (all >100bp non-coding sequence). PAR genes showing significiatnly higher divergence are shown.

Bianchi and Li 0.774 (± 0.064)) (Estimated in MEGA2 - Kumar et al. 2001). The silent sites for this gene are approaching, or have reached saturation and are clearly much more diverged than for the other sequences shared by the mouse and rat genomes.

To place the divergence values of these PAR regions in context, the spread of this measurement across the genome has been illustrated in the distribution shown in Figure II.2. As there are strong differences in the mutation rate between different regions of the genome (Lercher, Williams, and Hurst 2001), it is not clear exactly what distribution of sequence divergence is predicted between two species such as human and orangutan. Even if the mutation rate was even and all substitutions were neutral, the distribution would change shape with increasing divergence (due initially to edge effects starting from 0% divergence and then again as the nucleotide substitutions approach saturation around 75% divergence).

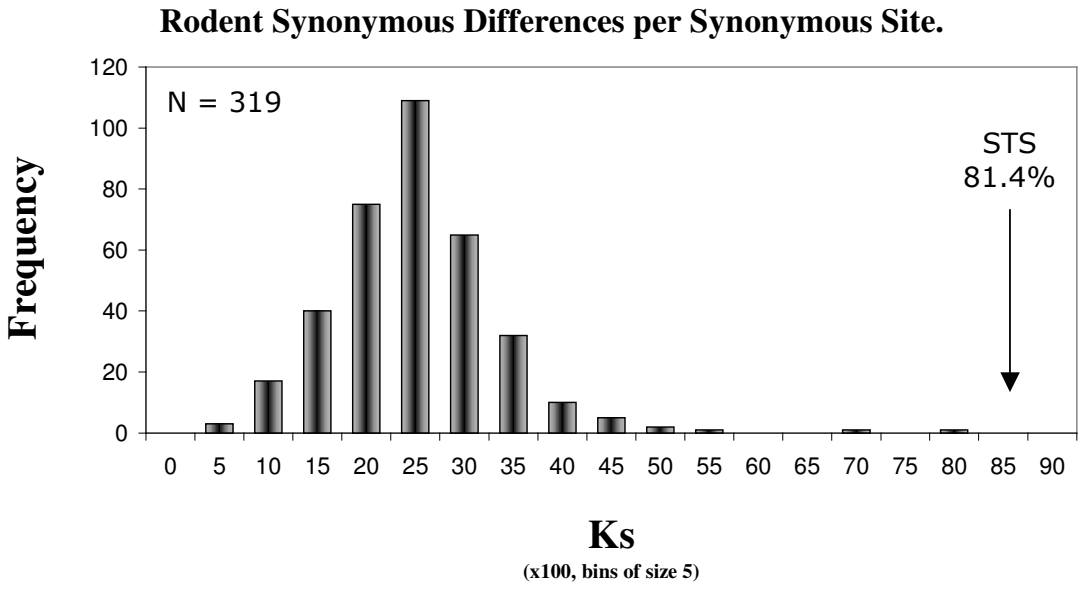


Figure II.3 The distribution of mouse-rat Ks values from Wolfe and Sharp (1993). The equivalent value for STS is 81.4%.

The distribution of orangutan-human divergences fits a normal distribution (X^2 = 3.409, d.f.=5, P=0.637, outlier classes combined to total 5+). The mean across all 51 divergence values was 2.50% (S.D.=0.96%), slightly lower than the average using a concatenation of all sites (2.85%). This may be caused by a length effect with shorter sequences having more extreme divergence values both above and below the mean. But we tried to control for this by removing sequences with less than 100 non-coding nucleotides and such an effect would mostly affect the variance, not the mean. More likely, the shorter non-coding sequences contain a greater proportion of regulatory elements as they include some promoters and 5' and 3' un-translated regions which are less neutral than nucleotides in the middle of large introns or intergenic regions. The mean divergence across the 17 alignments with 1000bp or greater, was 2.90% (S.D.= 0.72%). It was reasonable to use all 51 sequence pairs because they represent near genic sequences similar to those sequenced in the PAR.

Table II.3 - Expected number of substitutions for each PAR gene following a Normal distribution with mean divergence =2.50% and standard deviation 0.9% from the distribution in Figure II.4.

Gene	Aligned length (non-coding)	Expected number of substitutions	Human-Orang substitutions (X)	P(X or greater)
PPP2R3LB	854	25.62	72	<0.001
SHOX	2157	64.71	120	<0.001
ASMT	722	21.66	51	<0.001
XG (+1 -2)	956	28.68	30	0.252
XG (PABxy)	1534	46.02	51	0.194
XG (PABx)	507	15.21	14	0.392
SYBL1	2200	66	60	0.406

The probabilities associated with obtaining each of the PAR sequence divergence values from this normal distribution are given in Table II.3. This is calculated using the mean and standard deviation of the distribution in Figure II.2. The results are qualitatively the same if the mean of the distribution is increased to 3.0% (in line with higher published estimates of human-orangutan divergence (Chen and Li 2001) and the standard deviation kept the same.

II.4 Discussion

We have shown that several newly sequenced near-genic regions from the orangutan PAR exhibit much higher neutral divergence from human than is normally found between these two genomes. A single gene from the rodent pseudo-autosomal region displays the same attribute when divergence is calculated using the rate of synonymous coding substitutions per synonymous site (K_s). These results agree with the divergence difference seen by Perry and Ashworth (1999) for Fxy, the gene which spans the PAR boundary of *Mus domesticus* but is unique to the X chromosome in *M. spretus*. If divergence at these silent sites is a fair estimator of the mutation rate, then such unusually high divergence in a region known to experience a storm of recombination is highly suggestive that recombination is a source of mutations.

As the sex chromosomes differ in their susceptibility to several evolutionary processes (see Chapter I), there may be other mechanisms causing an unusually high substitution rate in the PAR. This result is not however, due to the relative amount of time spent in males versus females. Being passed solely in the male line, the Y chromosome is exposed to more mitotic cell divisions than any other chromosome (in the production of sperm). This is the male driven evolution effect (Haldane 1947) and mutations caused by replication may be partly responsible for differences in the rate of evolution of X and Y linked genes. In this instance however, we can rule out such an effect because PAR linked

genes move freely between the sex chromosomes and spend equal time in males and females.

Nor is this result due to the Y chromosome having a high mutation rate *per se*. Regional genome differences may exist in the rates of evolution (Lercher, Williams, and Hurst 2001) and the Y chromosome in it's entirety may be a hotspot for mutation. This would cause Y linked genes to have higher divergence than the genome in general. However, the same study by Lercher, Williams et al. (2001) found that the rate on the X chromosome was significantly lower than the average for the autosomes (though so were some of the autosomes, but X was lowest). As the PAR genes spend only a quarter of their time on the Y chromosome (with equal sex ratios, there are three X chromosomes to every Y in the population), we would expect the rate of mutation to be closer to that of the X specific genes than the NRY (Non-Recombining Y) genes. However, looking at human-orangutan divergence for the NRY genes in Table II.4, it is clear that the PAR divergence values are at least as great as the rest of the Y chromosome and possibly greater.

Table II.4 – Human-Orangutan divergence of four Y specific genes.

Y Region	Length	Divergence	Reference
ZFY intron	~1000bp	3.9%	(Shimmin, Chang, and Li 1993)
TSPY intron	755bp	8.2%	(Kim and Takenaka 1996)
SMCY intron	4758bp	5.42%	(Shen et al. 2000)
AMELY	1198bp	4.93%	(Huang et al. 1997)

The high divergence seen throughout the PAR is unlikely to be a telomeric effect. Baird and Royle, looking first specifically at the most distal part of

the PAR, near the X/Y telomere (Baird and Royle 1997), and then more generally at this subtelomeric region on other human chromosomes (Baird et al. 2000), discovered distinct haplotypes featuring high sequence polymorphism and strong linkage disequilibrium between them. These distinct haplotypes however only extend at most 2kb distally from the telomere boundary and cannot explain the high divergence values spanning the more than 2Mb of the PAR.

Several groups have recently investigated a genome-wide link between divergence and recombination. Previously, Nachman et al.(1998) (See also Nachman 2001) found a correlation between recombination and heterozygosity for non-coding sequence associated with a range of (mostly X-linked) genes. However, they did not find that the correlation was also between divergence and recombination rate and therefore did not support the recombination mutagenesis hypothesis. Instead, they found that their data were better explained by selective sweeps controlling diversity in those regions.

Using the set of human-orangutan divergence values collated for this study, we looked for a correlation with the human recombinational map. Figure II.6 shows that we found none.

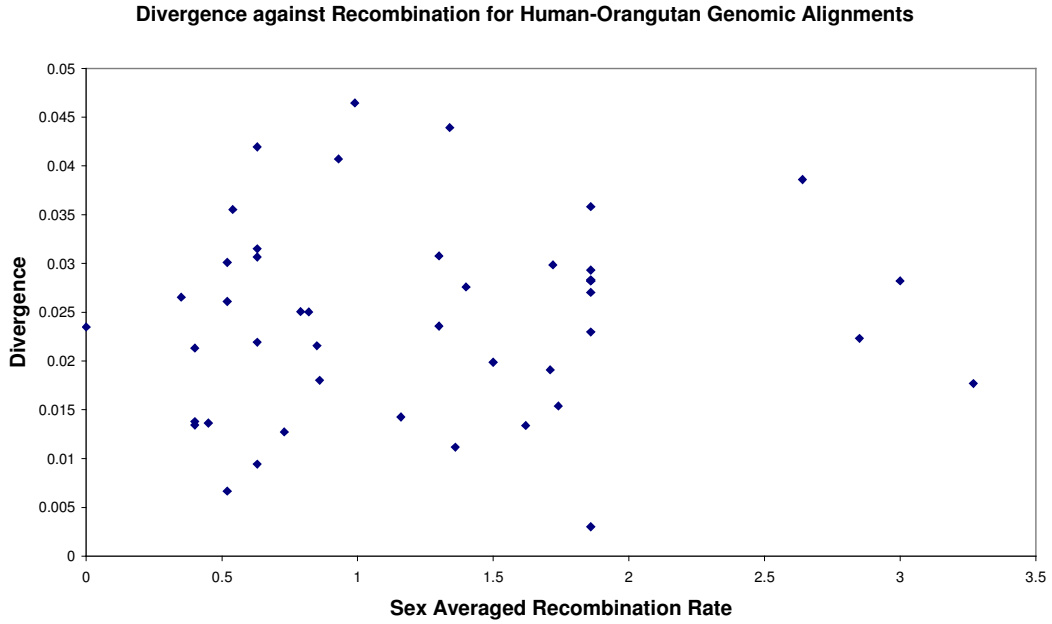


Figure II.4 – Human-orangutan divergence plotted against the sex-averaged recombination rate in humans.

Recently however, Lercher and Hurst (2002) found a weak (R^2 <0.1), but very significant correlation between SNP diversity and recombination rate throughout the genome. As less than 5 % of these SNPs are within coding regions (based on the proportion of the genome that is coding), and therefore most avoid selective sweeps without excessive recombination, this is evidence still for a mutagenic role for recombination. In addition Hellmann et al. (2003) were able to compare 115 new BAC sequences from baboon (an Old World Monkey) each over 10kb in length with a newer, more detailed recombination map of the human genome (Kong et al. 2002). They found a stronger significant correlation between recombination and divergence (Pearson's correlation coefficient =0.264). Together, these results suggests a weak mutagenic effect for

recombination throughout the genome. Around coding regions, selective sweeps may override neutral fixation.

The SHOX sequence used in this study was made up of three discrete non-coding regions: SHOX(+1/-12), SHOX(+15/-17) and SHOX(+7/-9) (numbers correspond to primer pairs). Interestingly, the three regions differ in their estimate of divergence (5.7%, 4.3% and 7.0%, respectively) and, whilst they are all greater than the genomic average, the region surrounding a known recombinational hotspot (SHOX +15/-17) (May et al. 2002), has the lower divergence. If recombinational hotspots move with time, the SHOX hotspot could be at its present location because of low diversity and therefore high homology between heterozygotes. Over time, this homology would decline and the hotspot would move to another location near SHOX. The repetition of this process could eventually lead to the overall increase in divergence between human and orangutan. This mechanism is compatible with the clustered distribution of recombination across the genome (McVean et al. 2004). The high divergence in the PAR is then a result of forced recombination in one region for an extended period of time. As long as new polymorphisms are not so common that homology between heterozygotes breaks down entirely, recombination will move up and down the PAR.

II.4.1 Conclusions

We have shown that non-coding, near genic regions of the orangutan pseudoautosomal region are significantly more diverged from human than

a representative distribution of comparisons from around the genome. This result adds to a previous finding for a mouse PAR gene and is augmented by our study of another rodent gene, STS and its comparison with like values from around the genome. Together, these results strongly support a link between the rates of recombination and divergence and, by proxy, the rate of mutation.

II.4.2 Epilogue

Another study using the introns of the PAB-spanning gene, XG, also found that divergence in this region was not significantly greater than elsewhere in the genome and concluded that any mutagenic effect of recombination is at best, weak (Yi et al. 2004). However, further work by D. Filatov using more regions throughout the PAR, has discovered a gradient of substitution rate increasing away from the PAB (Filatov 2004). This would be consistent with interference of chiasma formation near the PAB.

II.5 Publication

The work presented in this chapter contributed to the following publication:-

High mutation rates in human and ape pseudoautosomal genes

Dmitry A. Filatov, Dave T. Gerrard

Gene 317 (2003) 67-77

The work outlined in the methods and results section of this thesis chapter was my contribution to this paper. D. Filatov wrote the majority of the paper with sections on the building of distributions of K_s scores written by myself. I proofread the entire document and made numerous small changes to improve grammar and clarity.

List of sections in this paper to which my work made a significant contribution:-

2.2 Building a distribution of human-orangutan divergence

2.3 Mouse-rat PAR divergence

3.1 Human/Orangutan divergence

3.2 Mouse/Rat divergence

4.1 Substitution rates in the PARs

4.2 Causes of elevated substitution rate in pseudoautosomal regions

III POSITIVE AND NEGATIVE SELECTION ON MAMMALIAN Y CHROMOSOMES

Dave T. Gerrard & Dmitry A. Filatov

School of Biosciences,

The University of Birmingham,

Edgbaston, Birmingham, B15 2TT, UK

Running title: Selection on the Y

Keywords: adaptive evolution; mammalian sex chromosomes;

Submitted to *Molecular Biology and Evolution*, November 2004

(small formatting changes have been made for presentation in this thesis)

Abstract

Y chromosomes are genetically degenerate in most organisms studied. The loss of genes from Y chromosomes is thought to be due to the inefficiency of purifying selection in non-recombining regions, which leads to the accumulation of deleterious mutations via the processes of hitchhiking, background selection and Muller's ratchet. As the severity of these processes depends on the number of functional genes linked together on the non-recombining Y, it is not clear whether these processes are still at work on the old, gene poor mammalian Y chromosomes. If purifying selection is indeed less efficient in the Y-linked, compared to the X-linked genes, deleterious non-synonymous substitutions are expected to accumulate faster on the Y chromosome. However, positive selection in Y-linked genes could also increase the rate of amino acid changing substitutions. Thus, the previous reports of an elevated non-synonymous substitution rate in Y-linked genes are still open to interpretation. Here we report evidence for positive selection in two out of three studied mammalian Y-linked genes, suggesting that adaptive Darwinian evolution may be quite common on mammalian Y chromosomes. Taking positive selection into account we demonstrate that purifying selection is less efficient in mammalian Y-linked genes compared to their X-linked homologues, suggesting that these genes continue to degenerate.

III.1 Introduction

Old Y chromosomes, such as those in mammals, birds and in *Drosophila* are genetically degenerate (Bull 1983), containing few functional genes (e.g. Skaletsky et al. 2003). It is thought that Y-linked genes accumulate deleterious mutations due to the reduced efficacy of purifying selection in the non-recombining regions (reviewed in Charlesworth and Charlesworth 2000). Deleterious mutations may be carried to fixation by linked advantageous mutations ("selective sweeps") (Rice 1987). Additionally, the selective elimination of deleterious mutations, causing "background selection" (Charlesworth, Morgan, and Charlesworth 1992) could accelerate the stochastic fixation of mildly detrimental mutations (Charlesworth and Charlesworth 2000). Furthermore, selective sweeps and background selection reduce the effective population size (and therefore variability) of genes in evolving Y chromosomes, allowing the operation of "Muller's ratchet" (the stochastic loss of chromosomes with the fewest mutations)(Charlesworth and Charlesworth 2000; Gordo and Charlesworth 2000). Both reduced genetic diversity and the accumulation of deleterious mutations were indeed reported for the young (10-20 million years old) Y chromosomes of the white campion *Silene latifolia* (Guttman and Charlesworth 1998; Filatov et al. 2001; Filatov and Charlesworth 2002; Matsunaga et al. 2003, Filatov, submitted), and for the neo-Y chromosomes of *Drosophila miranda* (Bachtrog and Charlesworth 2002; Bachtrog 2003), supporting these hypotheses.

With time, the accumulation of deleterious mutations, gene loss and inability of Y-linked genes to adapt to a changing environment are expected to lead to genetically degenerate Y chromosomes, similar to the mammalian Y.

But what is the further fate of the gene poor mammalian Y chromosome? Will it degenerate further until it is no longer required, perhaps leading to X0 males as in the rodent Mole vole, *Ellobius lutescens* (Fredga 1988)? Or worse, it has been claimed that the loss of human Y-linked genes is inexorable and that it may lead to the extinction of the entire species because of the active role this chromosome plays in sex determination (Aitken and Marshall Graves 2002; Sykes 2003). Yet, the efficacy of each of the processes causing genetic degeneration depends on the number of functional genes linked together on the non-recombining Y chromosomes (Charlesworth, Morgan, and Charlesworth 1993; Peck 1994; Orr and Kim 1998; Charlesworth and Charlesworth 2000; McVean and Charlesworth 2000). With few functional genes surviving on the modern mammalian Y chromosomes (e.g. Skaletsky et al. 2003), the per-chromosome deleterious mutation rate may be too low for background selection to operate (Charlesworth, Morgan, and Charlesworth 1993) and adaptive mutations may also be too rare for selective sweeps to play a major role in the accumulation of deleterious mutations. In contrast to the drastic reduction of genetic diversity on the young *Silene latifolia* Y chromosomes ((Filatov et al. 2000; Filatov et al. 2001; Matsunaga et al. 2003) and *Drosophila miranda* neo-Y (Bachtrog and Charlesworth 2002; Bachtrog 2003), DNA diversity is only slightly lower on the human Y than on the X (Nachman 1998; Shen et al. 2000). This is consistent with the hypothesis that, after 240-320 million years of degeneration (Lahn and Page 1999), the human Y contains too few genes for the degeneration processes to actively proceed further. To address the question of whether the degeneration of the old mammalian Y chromosomes still continues, we analysed the amino acid

and synonymous substitution rates in three pairs of mammalian Y- and X-linked genes.

Comparing homologous coding DNA sequences from several species, it is possible to infer selection from the ratio of divergence rates at nonsynonymous (K_a) and silent (K_s) sites. If purifying selection is eliminating most amino acid replacements before they fix in the population and become substitutions, then K_a/K_s will be much less than one, whilst positive selection at many codons should cause this ratio to exceed unity. If the efficacy of purifying selection on the Y chromosome is reduced, as suggested by theory (Felsenstein 1974; Rice 1996), then the K_a/K_s ratio should be higher for the Y-linked genes, compared to the X-linked homologues. Indeed, several studies have already reported an elevated number of amino acid substitutions in Y (or W) versus X (or Z) linked genes (Agulnik et al. 1998; Fridolfsson and Ellegren 2000; Wyckoff, Li, and Wu 2002; Bachtrog 2004). However, there remains the possibility that positive selection was partly responsible for the increased K_a/K_s on the Y relative to the X. Adaptive non-synonymous changes at a small number of sites would increase K_a/K_s but the gene-wide average ratio would be kept below unity by purifying selection operating on the majority of codons. This possibility has not been addressed in the previous comparisons of the K_a/K_s ratios among the X- and Y-linked genes (Agulnik et al. 1998; Fridolfsson and Ellegren 2000; Wyckoff, Li, and Wu 2002; Bachtrog 2004).

To estimate the relative impacts of adaptive and purifying selection in the homologous mammalian X- and Y-linked genes we sequenced partial coding regions of three pairs of sex-linked genes, *SMCX* / *SMCY*, *USP9X* / *USP9Y* and *UTX* / *UTY*, from up to twelve mammalian species. These sequences were analysed using maximum likelihood models allowing for variable selective

pressures across codons (Yang et al. 2000). Unexpectedly, we detected positive selection in two out of three studied mammalian Y-linked genes, *USP9Y* and *UTY*. This suggests that positive selection in Y-linked genes is not uncommon and should be accounted for when inferring patterns of selection from substitution rates in X- and Y- linked genes. Furthermore, after correcting for positive selection, we demonstrate that purifying selection is still less efficient on mammalian Y linked genes than their X linked homologues – a sign that our Y may still be degenerating.

III.2 Materials and Methods

III.2.1 DNA Samples and Extraction

Pre-extracted DNA was obtained from Corriel Cell Repository for the following species (genes successfully amplified and sequenced in parentheses): *Lemur catta*, Ring-tailed lemur (*SMCX*, *UTX*, *USP9X*); *Gorilla gorilla,* Gorilla (*SMCX*, *UTX*, *USP9X*); *Ateles geoffroyi*, Black spider monkey (*SMCX*, *SMCY*, *UTX*, *UTY*, *USP9X*, *USP9Y*); *Macaca fascicularis,* Crab-eating macaque (*SMCX*, *SMCY*, *UTX*, *UTY*, *USP9X*, *USP9Y*); *Pongo pygmaeus abelii,* Orangutan (Sumatran) (*UTX*, *UTY*, *USP9X*, *USP9Y*); *Pan troglodytes,* Chimpanzee (*UTY*); *Homo sapiens*, Human (*UTX*, *UTY*). In addition, DNA was extracted from tissue samples obtained from the Institute of Zoology, London for the following species: *Eulemur fulvus*, Brown lemur (*SMCY*, *UTY*, *USP9X*, *USP9Y*); *Cheirogaleus medius*, Fat tailed dwarf lemur (*UTX*); *Leontopithecus rosalia*, Golden lion tamarin (*UTX*, *USP9X*); *Presbytis entellus*, Hanuman langur (*UTX*); *Colobus sp.*, Black and white colobus monkey (*UTX*); *Hylobates lar*, Lar gibbon (*UTX*, *UTY*, *USP9X*, *USP9Y, SMCY*); *Gorilla gorilla*, Gorilla (western lowland) (*UTY*, *USP9Y, SMCY*). Human, chimpanzee and mouse sequences were obtained from GenBank unless specified otherwise above. Total genomic DNA was extracted using DNAzol reagent (Invitrogen).

III.2.2 PCR, Cloning & Sequencing

PCR was carried out in 25µl reactions using the following general recipe in which 'T_1' represents the primer specific annealing temperature and T_2 is T_1+ 2ºC: 95ºC for 2mins 30s; T_2ºC for 30C; 72ºC for 3mins 30s; 92ºC for 20s; T_1ºC for 30s; 72ºC for 3mins; 33 cycles of steps 4 – 6; 72ºC for 5mins; 7ºC hold. Table III.1 lists the primers used for PCR in each gene and also the internal primers where they were necessary. All primers were designed on GenBank human sequence using Vector NTI. Where PCR products were weak, or direct sequencing indicated multiple products, they were cloned into pCR plasmids using the TA cloning kit (Invitrogen). EcoR1 digestion was used to confirm the presence/absence of cloned inserts. Sequencing was carried out in 5µl reactions using 2µl of gel-extracted DNA for direct sequencing or 0.5µl of plasmid from cloned DNA. The sequencing used the BigDye v3.1 kit (ABI) on an ABI3700 capillary sequencer. All sequences were covered in forward and reverse directions. For each gene, a single, multi-species 'contig' was generated in Gap4 (Staden, Beal, and Bonfield 2000) using PreGap4 to feed in the raw ABI sequence files. Assembling one contig for all species automatically aligned the sequences and divergent indels and site substitutions were revealed early on and could be checked either by simply referring back to the trace or by re-sequencing a piece of sequence if necessary. Novel sequences have been submitted to Genbank under accession numbers AY591386-AY591426 & AY699605-AY699606.

Table III. 1 – PCR & Sequencing Primers

Gene	Primer Name	Primer Sequence (5'-3')
SMCX	SMCX_+5	CACTTGAGGCCATAATCCGTGAAG
SMCX	SMCX_+1	ATGGTGACCACTACCCCTGC
SMCX	SMCX_-6	TTTGTACAACCCCAGCTCCTTCTC
SMCX	SMCX_+12	GTCCACTGGCATCATCGAGC
SMCX	SMCX_-3	GGTAGTTAGGAGGCTCCTCAGGTC
SMCX	SMCX_+7	GGCGGAGGGTCTCAAGTTTG
SMCX	SMCX_-2	TCTGGTTCTCCTGGGTGCTG
SMCY	SMCY_+7	ACCCCTGTCTAGATGACTTGGAG
SMCY	SMCY_-12	TCTCTGAGGTCCTGTGCAGACAG
SMCY	SMCY_+16	AGACAATCCTAGCCTTGCTGG
SMCY	SMCY_-8	AGACACTGAAGGGCCTCACC
SMCY	SMCY_+13	AGGTGAGCTTTCCAGGCCAG
SMCY	SMCY_-6	TGCGGCAACAGTGAGGACAG
SMCY	SMCY_-2	AGCCTGAATCCTTGTGTTGT
USP9X	USP9X_+1	TGATACCGTAAAGCGCTTGC
USP9X	USP9X_+3	CGTATGTCATACAGGCGCCA
USP9X	USP9X_-4	GGAGGGTTTAAGTAAACGCCAT
USP9X	USP9X_-2	TGGCAGAGCCACGGACTACT
USP9Y	USP9Y_+1	TTGCCTCGGGTTCTTGCTAT
USP9Y	USP9Y_+3	TGATGATGGAGATGTAACAGAATGC
USP9Y	USP9Y_-4	CTGATGGGGTCTTGCAATAGTTA
USP9Y	USP9Y_-2	TGGCAGGACCACGAACTATT
UTX	UTX_+1	GGTCAGAGTTCACATTCGGC
UTX	UTX_-2	AGAACTTCTGCTGAGCTGGG
UTY	UTY_+1	CAGAGTTCATGTTTGTCAGGACC
UTY	UTY_-2	GCATGCTTTCAGAACTTCTGTTG

III.2.3 Sequence analysis

Species consensus sequences were extracted from Gap4 with alignment gaps (indels) intact. In Proseq 2.9 (Filatov 2002), coding sequences were assigned based on the human intron/exon structure given in Genbank. The *USP9Y* sequence from Black spider monkey (*Ateles geoffroyi*) contained both a stop codon and, further downstream, a frameshift mutation.

As the region sequenced for this study began in exon 36 of 46 and 70% of the coding sequence lies before this point (based on human sequence), we cannot be sure whether *USP9Y* is non-functional or just curtailed in this species. These were the only stop codons and frameshifts detected in the study and this sequence was excluded from any analyses.

III.2.4 Maximum Likelihood Estimation and Likelihood Ratio Tests

The maximum likelihood estimates of various evolutionary models were calculated using the programme *codeml*, a constituent of the PAML package (Version 3.14 beta, Yang 1997)), and its notation is followed here: the K_a/K_s ratio is termed ω and model names are as in Yang et al. (2000) and Swanson et al. (2003). In what follows, when we refer to different codons within a gene, we mean different positions along the protein sequence and not every occurrence of a certain coding triplet.

Most codons in most genes are probably under purifying selection and have $\omega<1$, while some genes may have few codons under positive selection (with $\omega>1$). Using models allowing for several classes of codons with separate ω ratios (Yang et al. 2000; Swanson, Nielsen, and Yang 2003) it is possible to test for the variation of selective pressures across the codons and for the presence of codons under positive selection. The simplest model, M0, assumes a single ω ratio for all codons and all the branches of the phylogeny. Fitting a single ω ratio assumes that all the codons are under the same selective pressure and the estimated ω averages across the codons. As only few codons in a gene may be under positive selection, while the rest of the gene is under purifying selection, the average ω is likely to be below 1 despite the action of positive selection in the gene. Thus, estimating a single ω ratio is an extremely conservative approach to test for positive selection. Adding another class of codons with a separate ω_1, may better describe the distribution of ω across codons. Model M3 allows for two or more classes of codons with separate ω ratios (Yang et al. 2000). The programme *codeml* was used to calculate the likelihood of each model separately e.g. L(M0) and L(M3), and a likelihood ratio test (LRT) was used to test if the more general model M3 fits the data better than M0. Under the null hypothesis, that M3 does not improve the fit of the model to the data, compared to model M0, the log-likelihood ratio, $2\Delta L = 2[\ln L(M0)-\ln L(M3)]$ is χ^2 distributed with degrees of freedom equal to the difference in number of parameters between the two models (Yang et al. 2000).

The significantly better fit of M3, compared to M0 demonstrates heterogeneity of ω among codons within the coding region of a gene. To test

for the presence of a class of codons under positive selection (ω>1), we used models M14 & M15 (Swanson, Nielsen, and Yang 2003). M14 fits a β distribution, which may be uni-modal, bimodal or flat, in the interval $0 \leq \omega \leq 1$ but does not allow any class of codons to have $\omega > 1$. M15 features an equally limited distribution but allows for an additional class of codons with $\omega > 1$. For the M14/M15 test, the *P*-value of the LRT is half that obtained from the normal chi-square statistic with one degree of freedom (Swanson, Nielsen, and Yang 2003) .

The models above were devised to detect positive selection, but do not allow testing of whether purifying selection differs significantly among the lineages. To test whether the Y-linked genes are less constrained by purifying selection, compared to the X-linked homologues, we compared a null model with a single ω ratio across the tree to the model with separate ω_x and ω_y ratios for the X- and Y-linked branches, respectively. The two models were compared using a likelihood ratio test (LRT), assuming that under the null hypothesis of no difference between the two nested models the test statistic, $2\Delta L = 2[\ln L(\text{model\_1}) - \ln L(\text{model\_2})]$ is χ^2 distributed with degrees of freedom equal to the difference in number of parameters between the two models (Yang 1998). For this analysis we used a combined tree for all the X- and Y-linked sequences for every gene (e.g. as shown in Figure 2) to calculate the likelihood under the two models. This test assumes a single ω ratio across codons, hence its results are equivalent to the previous comparisons of the K_a/K_s ratios among the X- and Y-linked genes (Agulnik et al. 1998; Fridolfsson and Ellegren 2000; Wyckoff, Li, and Wu 2002; Bachtrog 2004), and suffer from the same problem – averaging across the codons may bias the ω ratio

upwards if some codons evolve under positive selection. Thus, a model with variable ω ratios across codons has to be used to account for the possibility of positive selection.

To test whether purifying selection is relaxed in the Y-linked, compared to the X-linked gene, taking into account variable selective pressures across the codons, one needs to use a model which allows for several classes of codons and lineage-specific ω ratios (e.g. ω_{0X}, ω_{0Y}, ω_{1X} and ω_{1Y}), the so called "branch-site model". Two branch-site models have been implemented in *codeml* (Yang and Nielsen 2002); however, the tests using these models appear to be too liberal and unreliable (Zhang 2004). Instead, we followed a parametric bootstrap approach to generate a distribution for the ω_{0X}, ω_{0Y} and the proportion of codons falling into the ω_0 class under the M3 model with two site classes. The bootstrapping was conducted using a Perl script, which generated 1000 pseudo-replicates bootstrapping across the codons. Each "column" of the dataset, containing the same codon position from all the sequences in the alignment was stored as a separate entity in an array. Each pseudo-replicate was generated by randomly selecting the "columns" of codons from this array until a new alignment was created matching the length of the original (Felsenstein 1988). As the bootstrap pseudoreplicates are non-independent, standard tests of significance (e.g. t-test) are not applicable; hence, the bootstrap distribution was represented graphically (Figure 3).

All the maximum-likelihood analyses described above are conditional on a topology of a phylogenetic tree. Initially topologies were calculated for each gene using both the neighbour joining (Tamura-Nei distance) and maximum

parsimony methods in MEGA2 (Kumar et al. 2001). Phylogenetic trees using combined X and Y sequence alignments were constructed using the neighbour joining method to test the phylogenetic relationships of the sequences and specifically to check for monophyletic origins of all X and all Y sequences. This may not have been the case if (i) the cessation of recombination between X and Y had occurred independently in different lineages, as it was demonstrated for bird sex chromosomes (Ellegren and Carmichael 2001), or (ii), intra-specific gene conversion had overwritten X with Y or vice versa as has happened at least twice to the *ZFY* gene during the radiation of cats (Pecon Slattery, Sanner-Wachter, and O'Brien 2000) and separately in the crab-eating fox, *Dusicyon thous* (Pamilo and Bianchi 1993). The first possibility seems unlikely for rodents and primates as the studied genes are thought to have become sex linked before the rodent/primate split (Lahn and Page 1999). Additionally, to test the sensitivity of the results of the likelihood ratio tests to changes in the tree topology, we re-ran the significant M14/M15 results using up to 47 alternative best trees generated in PAUP* (Swofford 1998).

III.2.5 Pairwise human-mouse X linked measurements

To place the substitution rates in the *UTX/Y*, *SMCX/Y* and *USP9X/Y* genes in context, we compared the pairwise ω estimates (between human and mouse) of nine ubiquitously expressed X linked genes possessing Y homologues with 121 other X linked genes. To do this, the map of human-mouse orthologous loci was cross-referenced with the list of mouse and human homologous

sequence pairs downloaded from LocusLink[ϕ] using locus id numbers common to both. This resulted in a list of Genbank mRNA accessions of human X-linked loci for which reciprocal best matches had been made with known and available rodent (mouse and rat) mRNAs. For each pair of mRNA accessions, the full Genbank entries were downloaded from the NCBI website. The following EMBOSS (Rice, Longden, and Bleasby 2000) programmes were used: Coderet – to extract and translate the coding nucleotide sequence; Needle – to align each pair of protein sequences; Tranalign – to align the nucleotide sequences based on the protein alignment. The PAML programme codeml (Yang 1997) was then used to estimate the pairwise ω for each nucleotide alignment using the method of Goldman and Yang (Goldman and Yang 1994). The sequence downloads, alignments and ω estimations were automated using a Perl script running on a LINUX platform. Handling and conversion of sequences and alignments were made possible with BioPerl[Ψ] extension modules.

[ϕ] ftp://ftp.ncbi.nih.gov/refseq/LocusLink/
[Ψ] http://www.bioperl.org

III.3 Results

To estimate the relative impacts of adaptive and purifying selection in the homologous X- and Y-linked genes we sequenced partial coding regions of three pairs of sex-linked genes, *SMCX* / *SMCY*, *USP9X* / *USP9Y* and *UTX* / *UTY*, from up to twelve mammalian species (Table III.2, and Methods section).

Table III.2 – Sequences obtained for each gene.
[a] The number of species studied for each gene. See experimental procedures for the list of species. [b] Based on the exons annotated on human sequence in GenBank [c] The number of codons is one-third the length of coding DNA sequence. [d] Based on length of human coding sequence annotated in GenBank

Gene	N[a]	Exons[b]	Codons [c]	% of cDNA[d]
X linked				
SMCX	6	21-27	457	30%
USP9X	10	36-37	235	9%
UTX	12	17	223	19%
Y linked				
SMCY	9	21-27	447	29%
USP9Y	7	36-37	242	9%
UTY	9	17	211	18%

III.3.1 Phylogeny of new X and Y sequences

Before carrying out any tests of evolutionary models, we examined the phylogenetic relationships of the sequences. The trees from *SMCX/Y* followed the standard primate phylogeny shown in Figure III.1 (Purvis 1995) irrespective of the method used. However, whilst they still separated into X and Y clades, the trees from *USP9X/Y* and *UTX/Y* were not entirely consistent

with the same traditional phylogeny and also differed depending on the tree-building method used (Neighbour Joining with nucleotide or amino acid sequences and Maximum Parsimony with nucleotide sequences). Figure III.2 shows the nucleotide neighbour joining tree using all sites in the *USP9X/Y* alignment. The positions of the *Pongo* and *Hylobates* sequences within both X and Y halves of the tree were not as expected: *Pongo* clustered with *Hylobates* for the *USP9Y*, while *Hylobates* was closer to *Homo* and *Gorilla* than *Pongo* using the *USP9X* gene. The *SMCX/Y* sequences were substantially longer than those of the other genes (see Table 2) and we reasoned that the Y linked genes at least, must have had the same phylogeny, we accepted the traditional primate phylogeny over any other to use in our analyses. Using the alternative phylogenies made no difference to the results of the likelihood ratio tests (see below).

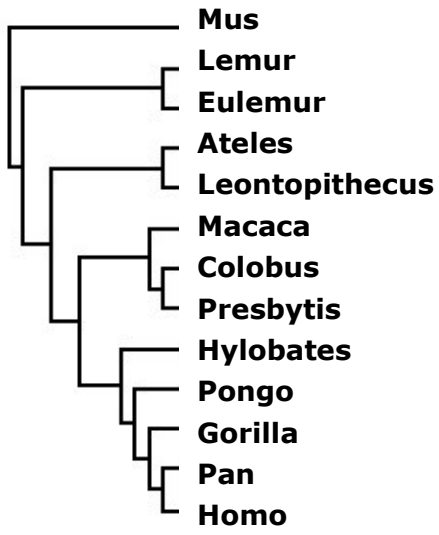


Figure III.1 - Phylogenetic relationship of the genera used in this study. Based on Purvis, 1995.

III.3.2 Testing for heterogeneity in Ka/Ks (ω) among codons

Datasets for each of the six genes were analysed separately using maximum likelihood models allowing for variable selective pressures across codons (Yang et al. 2000). The models allow the estimation of both the K_a/K_s ratio (ω, using the notation of the programme PAML (Yang 1997)) for several discrete classes of codons, and the proportion of codons within each class. The nested models with increasing numbers of parameters can be compared by

the likelihood ratio test (LRT), providing a test of whether a more parameter rich model fits the data significantly better (see methods).

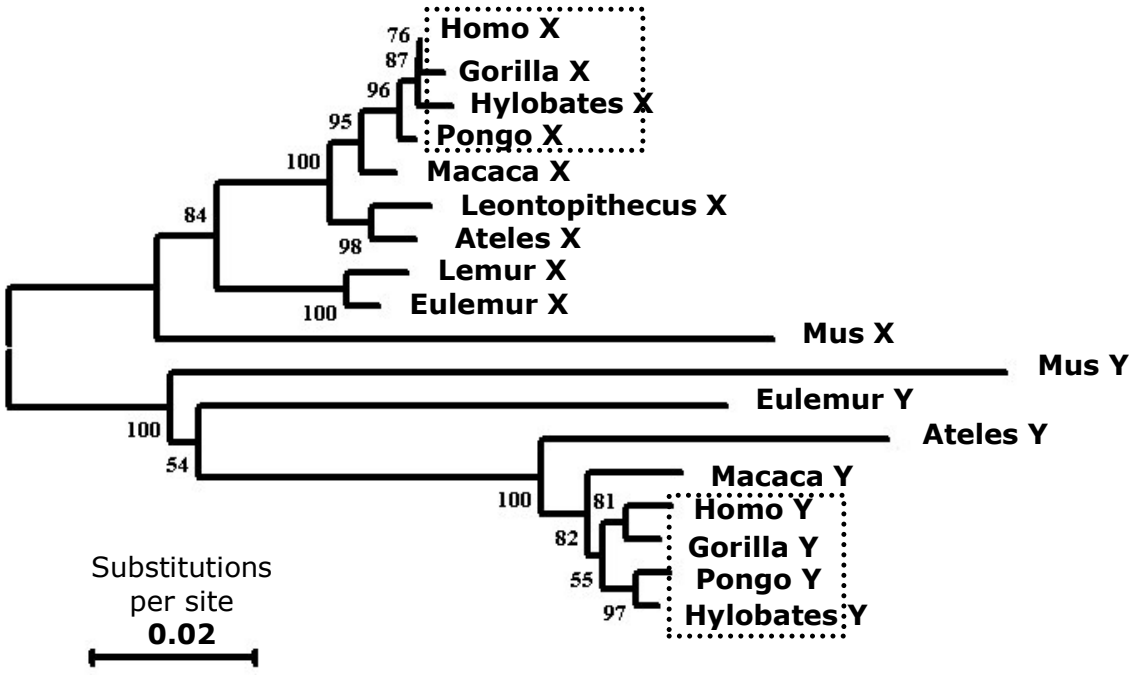


Figure III.2 – The neighbour joining tree (Tamura-Nei distance) of nucleotide sequences for *USP9X* and *USP9Y* sequences combined. The topologies of the ape clades (boxed) in both X and Y halves of the tree do not agree with the traditional phylogeny.

The analysis using model M0 (Yang et al. 2000) is similar to that from the pairwise comparisons conducted by others (Agulnik et al. 1998; Fridolfsson and Ellegren 2000; Wyckoff, Li, and Wu 2002; Bachtrog 2004), as it allowed for just a single class of codons. Consistent with previous results, the estimated ω ratio for the Y-linked genes is greater than that for their X-linked homologues (Table III.3). The significance of this difference can be shown on a combined tree of X and Y sequences using a model allowing for separate ω ratios for X- and Y-linked genes and testing this model against the model with a single ω ratio for all branches. For all 3 pairs of genes, the LRT showed that

separate ω for X and Y branches results in a significantly better fit of the model to the data (P<0.0001, degrees of freedom (d.o.f.) = 1, Table III.4).

Table III.3 - Likelihood ratio tests of models M0 vs. M3
[a] lnL represents the log-Likelihood [b] ω represents the codon-specific K_a/K_s estimate from PAML [c] 2ΔL represents double the difference in log likelihoods between the two models. [d] degrees of freedom = 2.

	Model M0		**Model M3** with two classes of codons				**M3 vs M0**	
Gene	**lnL [a]**	**ω^b**	**lnL**	**ω_0**	**ω_1**	**% with ω_1**	**2ΔL[c]**	**P-value[d]**
X-linked								
SMCX	-3240.7	0.149	-3206.0	0.000	1.060	19.3%	**69.385**	**<0.001**
USP9X	-1590.0	0.089	-1588.4	0.076	1.582	1.0%	3.117	0.210
UTX	-1237.3	0.270	-1232.4	0.014	1.580	16.5%	**9.897**	**0.042**
Y-linked								
SMCY	-4389.2	0.286	-4357.0	0.140	1.201	19.0%	**64.361**	**<0.001**
USP9Y	-1750.5	0.207	-1732.6	0.090	1.387	15.0%	**35.810**	**<0.001**
UTY	-1992.8	0.962	-1987.4	0.512	2.120	35.8%	**10.744**	**0.030**

Model M3 (Yang et al. 2000) takes into account the variable selective pressures across the codons by allowing for several classes of codons, with ω ratios estimated separately for each class. Model M3 with two classes of codons fits the data significantly better than the model M0 for all the genes except *USP9X* (Table III.3), demonstrating that selective pressure significantly varies across the codons in five out of six genes studied. The addition of extra classes of codons to model M3 did not improve the fit of the model to the data for any of the genes (data not shown), suggesting that two classes of codons are sufficient to accommodate the variation in selective regimes among the codons.

Table III.4 – Likelihood ratio tests of X and Y branch specific models on a combined tree

Genes (# of X/Y)	# of Codons	Separate rates	lnL	Branch ωs		
				ω_X	ω_0	ω_Y
UTX/*Y* 12/9	207	ω_0	-2537.246		0.743	
		ω_Y, ω_0	-2529.545**		0.383	1.063
		ω_X, ω_0	-2527.548**	0.243	0.965	
USP9X/*Y* 10/7	233	ω_0	-2699.612		0.167	
		ω_Y, ω_0	-2687.061**		0.078	0.272
		ω_X, ω_0	-2685.333**	0.065	0.261	
SMCX/*Y* 6/9	435	ω_0	-5811.661		0.228	
		ω_Y, ω_0	-5802.744**		0.146	0.287
		ω_X, ω_0	-5800.538**	0.134	0.289	

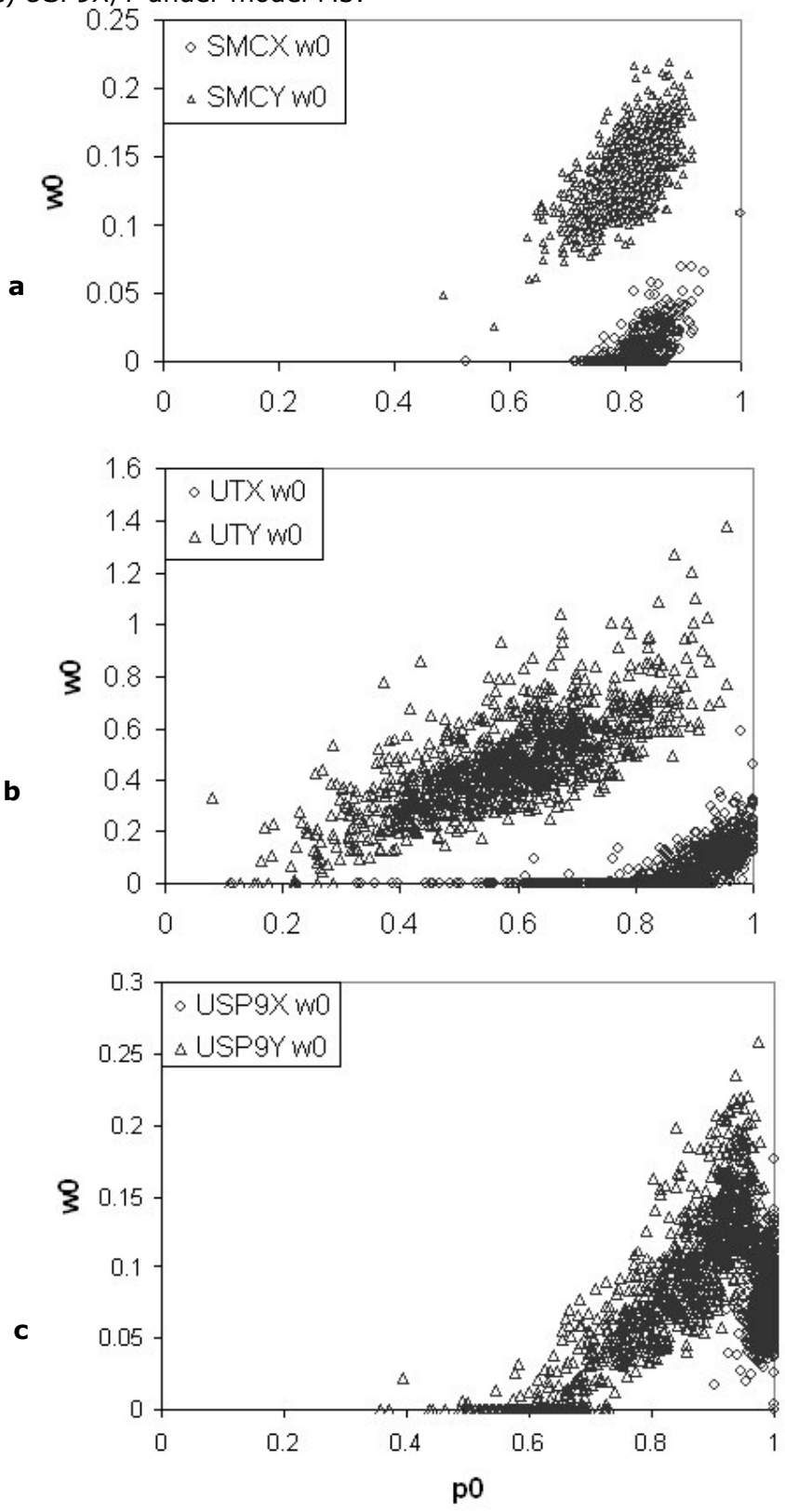
For each combined X and Y tree, three models were fitted to the data. The first model gave one ω ratio to all branches on the phylogeny (equivalent to Model 0 in previous models). The second and third models allowed two separate ω ratios: one for the Y, or the X specific branches respectively, with the rest of the tree having ω_0. In all cases a model with two different ω ratios within the tree fit the data better than with one overall ω_0. ** LRT between one ω for all branches and two omegas significant at $P<0.0001$, degrees of freedom = 1

Under model M3 with two classes of codons, the majority of the codons fall into the class with the lower ω_0 ratio (Table III.3), reflecting the fact that most codons in these genes are under purifying selection. Interestingly, the proportion of codons falling into the class with ω_0 is substantially lower in *USP9Y* (85.0%) and *UTY* (64.2%), compared to *USP9X* (99.0%) and *UTX* (83.5%). This reduction in the number of codons under purifying selection in *USP9Y* and *UTY*, compared to the X-linked homologues can be either due to reduced efficacy of purifying selection in the non-recombining Y-linked genes, or due to positive selection at some codons in the *USP9Y* and *UTY* genes. The

estimated ω_0 for this class of codons is higher for all the Y-linked genes, compared to the X-linked homologues (Table III.3). This is consistent with the results of the previous studies which reported an elevated number of amino acid substitutions in Y (or W) versus X (or Z) linked genes (Agulnik et al. 1998; Fridolfsson and Ellegren 2000; Wyckoff, Li, and Wu 2002; Bachtrog 2004), and with the theoretical prediction of a reduced efficacy of purifying selection on the non-recombining Y chromosomes (Charlesworth and Charlesworth 2000), suggesting that mammalian Y-linked genes continue to accumulate deleterious mutations at a greater rate than those on the X.

To test whether the Model M3 ω_0 values for the Y-linked genes are significantly higher than for the X-linked homologues, we followed a parametric bootstrap approach. For each gene we bootstrapped across codons to generate distributions of ω_0 and the proportion of codons falling into the ω_0 class, p_0 (Figure III.3). Assuming independence of bootstrap replicates, t-tests showed that ω_{0Y} is significantly greater than ω_{0X} for all three pairs of genes (P<0.001). However, the non-independence of the pseudo-replicates may make such a test too liberal. For *SMCX*/*Y* (Figure III.3a) and *UTX*/*Y* (Figure III.3b) the 'clouds' of values for the X- and Y-linked genes are clearly distinct, suggesting that the increase in ω_{0Y} over ω_{0X} is significant for these genes. On the other hand, for *USP9X*/*Y* there is considerable overlap of ω_{0Y} and ω_{0X} distributions (Figure III.3c), suggesting that ω_{0Y} is not significantly greater than ω_{0X} for this pair of genes.

Figure III.3 – Plots of ω_0 against the proportion of codons falling into the ω_0 class, p_0, from the 1000 bootstrap replicates for each of (a) *SMCX*/*Y*, (b) *UTX*/*Y* and (c) *USP9X*/*Y* under model M3.



III.3.3 Testing for positive selection at some codons

For the second class of codons, the ω_1 ratio exceeds unity for all the X and Y-linked genes (Table III.3), suggesting that some codons in these genes may be under positive selection. To test whether adaptive selection is indeed acting in all, or some of these genes, we have to demonstrate that the ω_1 ratios are significantly greater than unity (i.e. are non-neutral) for this class of codons. This can be done by comparing two nested models, one allowing for a class of codons free to take any $\omega>1$ and the other with this class having fixed $\omega = 1$. Models M14 and M15 (Swanson, Nielsen, and Yang 2003) implemented in PAML allow the ω ratios below 1 to be distributed according to a beta distribution and thus may better account for the variation in purifying selection across the codons, compared to models M0 and M3. To accommodate the codons with ω ratios above 1.0, these models have a further class (ω_1), that, in M15 is free to take any K_a/K_s value above 1.0, and is fixed at $K_a/K_s = 1$ for model M14 (Figure III.3).

The comparison of M14 and M15 by the LRTs demonstrated that model M15 fits the data significantly ($P < 0.05$, d.o.f. = 1) better than M14 for *UTY* and *USP9Y*, but not for *SMCY*, or any of the three X-linked genes (Table III.5), suggesting the presence of codons under positive selection in *UTY* and *USP9Y* genes.

To assess whether the topology of the tree that we chose affected the results of the likelihood ratio tests, we ran the M14/M15 tests for *UTY* and *USP9Y* using different tree topologies. We used PAUP 4.0 (Swofford 1998) to

generate alternative trees close to the best tree topology. The trees were allowed to contain multi-furcating nodes but were all rooted using the mouse sequence as an outgroup. For *UTY*, 45 alternative trees were used. Twelve of these trees were no worse than the best tree (P< 0.05, KH-test (Kishino and Hasegawa 1989)). None of the 45 trees failed to show the significantly better fit of Model 15 compared to Model 14 in the log likelihood ratio test after these models were run for each tree. For *USP9Y*, 47 similar trees were used with 25 as good as the best tree (KH-test, P<0.05). Again, none of the tree topologies affected the result of the M14/M15 ratio test and the model including adaptive selection better explained the data for all trees.

Table III.5 - Likelihood ratio tests of models M14 vs M15.
[a] degrees of freedom = 1

	M14		M15			LRT	
Gene	**lnL**	**% with ω = 1.0**	**lnL**	**ω_1**	**% with ω_1**	**2ΔL M15 vs. M14**	***P*-value[a]**
X-linked							
SMCX	-3206.1	19.7%	-3206.0	1.060	19.3%	0.095	0.379
USP9X	-1595.3	16.4%	-1595.3	1.000	16.4%	0.000	0.500
UTX	-1232.7	24.0%	-1232.4	1.646	14.9%	0.609	0.218
Y-linked							
SMCY	-4357.4	22.8%	-4357.0	1.211	18.7%	0.7788	0.189
USP9Y	-1732.7	18.7%	-1730.7	4.089	3.7%	**3.859**	**0.025**
UTY	-1989.8	81.2%	-1987.4	1.988	38.9%	**4.777**	**0.014**

III.4 Discussion

The accuracy and power of the phylogenetic maximum likelihood analysis to detect positive selection has been extensively tested (Anisimova, Bielawski, and Yang 2001; Zhang 2004). Zhang (2004) demonstrated that branch-site likelihood methods (Yang and Nielsen 2002) are unreliable, as they detect positive selection more often than 5%, when applied to data simulated with an absence of selection. However, this does not apply to the codon-based likelihood analyses used in our paper. The site-branch methods are designed to infer selection at some of the sites along some of the branches in the phylogeny, and the authors are very cautious about these methods (Yang and Nielsen 2002). On the contrary, the method used in our paper tests for selection on individual codon sites over an entire phylogenetic tree (Yang et al. 2000). It was demonstrated that this approach is fairly conservative and reliable (Anisimova, Bielawski, and Yang 2001). Models M14 and M15 (which are modifications of model M8) are much more stringent, compared to M0/M3 (Anisimova, Bielawski, and Yang 2001). Indeed, using alternative tree topologies had no effect on the result of the M14/M15 ratio tests. Thus, the rejection of model M14 for *UTY* and *USP9Y* strongly suggests the presence of codons under positive selection in these genes.

The detection of positive selection for two of the three Y-linked genes studied suggests that adaptive selection may not be a rare phenomenon on the mammalian Y chromosome. Indeed, an analysis of the mammalian Y-linked *DAZ* gene family demonstrated that the high ω ratio in this gene is caused not by the absence of purifying selection, as thought before (Agulnik et al. 1998), but by adaptive selection at some of the codons, while most codons

are under purifying selection (Bielawski and Yang 2001). The *DAZ* gene family belongs to the class II of Y-linked genes, according to the classification proposed by Lahn *et al.* (2001), i.e. it is a Y-specific gene family with a male-specific function, which evolved on the Y chromosome. Such genes are probably more free to evolve new functions than the genes in this study which match their X homologues in the breadth of expression and, to an unknown extent, function. Our results provide evidence for positive selection in single-copy mammalian Y-linked genes.

The causes of the positive selection in *UTY* and *USP9Y* are not completely clear. *UTX*/*Y* and *USP9X*/*Y* belong to a small group of ubiquitously expressed housekeeping genes with active copies resident singly on X and Y. The genes in this group are under fairly strong purifying selection: the pairwise ω for the mouse / human divergence for nine X-linked genes (*USP9X, SMCX, UTX, UBE1X, ZFX, SOX3, DBX, RBMX, RPS4X*) which retained expressed Y-linked homologues are significantly lower, compared to 121 X-linked genes without active human Y-homologues (P = 0.022, Kruskal-Wallis test). Thus, it is possible that these genes retained active Y-linked homologues due to more stringent purifying selection, compared to X-linked genes that lost Y homologues. Positive selection in the *UTY* and *USP9Y*, but not in the X-linked copies of these genes may suggest that the X- and Y-linked copies are somewhat diverged in function. However, the X-copies of these genes are not dosage compensated (Lahn, Pearson, and Jegalian 2001), suggesting that both X- and Y-copies of these genes perform similar functions. Alternatively, the positive selection in the *UTY* and *USP9Y* genes may be for compensatory mutations, which maintain the function of the Y-linked genes despite the accumulation of deleterious mutations in the non-recombining genes.

Our finding that two out of three Y linked genes studied undergo positive Darwinian selection indicates that it may be fairly common on the Y chromosomes. Previously, the elevated ω ratios in the mammalian (or avian) single-copy Y-linked (or W-linked) genes, compared to the X-linked (or Z-linked) homologues, were interpreted as evidence for relaxation of purifying constraint on the Y chromosome (Agulnik et al. 1998; Fridolfsson and Ellegren 2000; Wyckoff, Li, and Wu 2002; Bachtrog 2004). Our results demonstrate that such interpretations have to be taken with caution because the situation may be complicated by the presence of adaptive selection in the Y-linked genes.

Acknowledgements This work was supported by a grant to D.A.F. from the BBSRC. DTG was supported by a PhD studentship from the BBSRC.

IV THE RATE OF EVOLUTIONARY CHANGE OF THE SEX LINKED AMELOGENINS IN PRIMATES

Dave T. Gerrard

School of Biosciences,

The University of Birmingham,

Edgbaston, Birmingham, B15 2TT, UK

IV.1 Introduction

The arrest of recombination between evolving X and Y chromosomes effectively duplicates the genes therein; except that each copy then begins to experience a new and different genomic environment. Whilst the X-linked copy suffers a modest reduction in population size (3/4 that of autosomal genes assuming an equal sex ratio)(Caballero 1995) it may still recombine when in females to generate new chromosome haplotypes. However, for 'alleles' trapped on the Y chromosome, the reduction in population size is greater and, without recombination, new haplotypes can only be formed by mutation or gene conversion. These differences significantly alter the mode and direction of the future evolution of each new copy. For the Y linked genes, this often means extinction.

IV.1.1 The PAR, the PAR boundary & its migration

Outside of the non-recombining region (NRY), the sex chromosomes pair and recombine normally in the pseudoautosomal region (PAR). At the interface of the NRY and the PAR, 2.6Mb from the Xp-telomere, is the pseudo-autosomal boundary (PAB). Interestingly, whilst this boundary has been constant for many millions of years (Ellis et al. 1990; Lahn and Page 1999), in the long history of mammalian sex chromosomes it has moved several times as, for whatever reason, the NRY has expanded, each time dividing a new region of the PAR into discreet X and Y locations. The result in humans has been four differently aged strata (Lahn and Page 1999),

within which, the surviving pairs of X and Y genes have been separated by an equal amount of time.

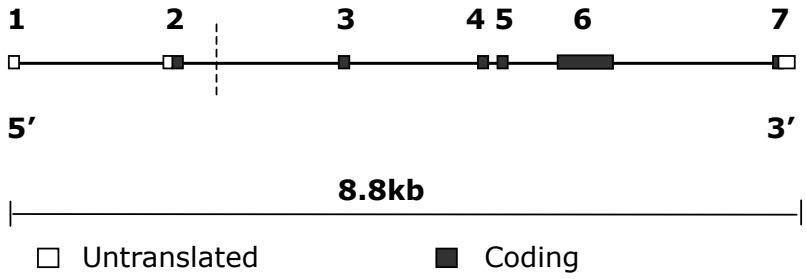
The degeneration of the NRY seems to have reached roughly the same stage in the three older strata, with very few X linked genes in these regions sharing a homologue on the Y (Lahn and Page 1999). Stratum 4 however, which arose amongst the monkeys after the divergence of Lemurs (60-90 MYA) but before the split of Old and New World monkeys (30-55 MYA)(Purvis 1995; Kumar and Hedges 1998; Glazko and Nei 2003), has relatively more pairs of X and Y linked gametologues (Lahn and Page 1999). However, of the seven pairs of X/Y gametologues identified by Lahn and Page (1999), the Y copy of five of these is a pseudogene. The % protein divergence is greater than the % nucleotide divergence and the pairwise Ka/Ks ratios between X and Y sequences are around 1.0, indicating neutrality.

IV.1.2 X linked Amelogenin lies on the site of an ancient PAB

On the human X chromosome, 11Mb from the p-telomere and at the boundary of strata 3 and 4, lies the amelogenin gene (AMELX), encoding a protein which helps lay down enamel in developing teeth. It is one of the few stratum 4 genes with a potentially functional Y homologue. The full 8.8kb transcript actually spans what would have once been the PAB and this now lies in the second intron (Iwase, Satta, and Takahata 2001; Iwase et al. 2003). A schematic of the exons and introns is given in Figure IV.1. The 5'-3' direction of the gene runs towards the telomere. Whilst

almost all the amelogenin coding exons (except 54bp of exon 2) lie in stratum 4 and hence were pseudoautosomal in the ancestor of Lemurs and Monkeys, the 5' UTR and any promoters already lay at this time on separate X and Y chromosomes (Iwase et al. 2003). The precise location of the old boundary was shown by measuring X-Y divergence along the gene in a range of eutherian mammals (Iwase et al. 2003). Divergence in the first two introns is within the range used to define stratum 3 (23-35%) (Lahn and Page 1999) whereas, after this point intronic divergence drops to around 10%, reflecting a more recent split of X and Y sequences. When the former were used to build a phylogenetic tree among several species, two distinct X and Y clades formed showing that the X-Y divergence pre-dated the diversification of the eutherian mammals. However, the more distal sequence gave a phylogenetic tree in which the X and Y sequences were more closely related to each other within some taxonomic groups than to X and Y sequences in other groups (Iwase et al. 2003). The cessation of recombination leading to the formation of stratum 4 had occurred at different times in different mammalian lineages.

Figure IV.1 - Exon pattern of human Amelogenin X. The position of the old PAR boundary is marked with a dashed line.



Three alternatively spliced isoforms of human AMELX have been described (NCBI). The most common excludes exon 4 and more rarely, there are

isoforms containing either all coding exons or missing both exons 4 and 5. It is believed that the splicing of AMELY is similar (Salido et al. 1992).

IV.1.3 The Function of amelogenin

The function of amelogenin has been well described (Lagerstrom et al. 1991; Salido et al. 1992; Alvesalo 1997; Mathur and Polly 2000; Hart et al. 2002). It is responsible for guiding the laying down of tooth enamel in growing teeth (Lagerstrom et al. 1991; Salido et al. 1992). The gene's expression seems to be limited to cells performing this role, suggesting a single function for amelogenin. As such, this gene differs from other functional genes shared by X and Y which are typically widely expressed. Additionally, one study of AMELY expression (in a 24 week old male foetus) found that the expression of this copy was only around 10% of the expression from the X chromosome (Salido et al. 1992).

The relative importance of X and Y copies of this gene have been examined largely in sex chromosome aneuploids. The phenotype of '45, X' humans feature a reduced enamel thickness on all teeth. In both '47, XXX' and '47, XXY' aneuploids, the enamel is significantly thicker than normal individuals (Alvesalo 1997). This suggests that the extra copy of AMELX is not inactivated, as is the case for most X-linked genes in this situation. Furthermore, '47, XYY' aneuploid males also have a similar increase in enamel thickness (Alvesalo 1997), indicating a similar scale of effect from X and Y copies. Whilst there are other differences, especially in tooth size, displayed by XYY individuals, these are probably due to another Y linked

gene TSY, located on the opposite arm to AMELY (Alvesalo and Portin 1980). In each aneuploid case, the effect of the lack of one chromosome or the addition of an extra X or Y seems to alter the enamel thickness by a metrically similar amount.

Whilst aneuploidy amounts to the deletion or duplication of the entire amelogenin locus, some different phenotypes have been described which arose from small mutations to the coding sequence (reviewed in Hart et al. 2002). Interestingly, family studies have shown that some such X-linked mutations have different effects in males and females. Typically, males have incomplete enamel formation, which is more susceptible to wear, while females have more normal enamel thickness, but each tooth features multiple vertical grooves (Hart et al. 2002).

IV.1.4 Amelogenin and comparisons between species

The association of amelogenin strictly with dentition is also supported by the presence, or absence, of its homologues in various non-mammalian vertebrates. Using blotting experiments, Girondot and Sire (1998) detected amelogenin homologues in *Danio rerio* (a fish) and *Varanus niloticus* (a lizard) but not in birds or turtles, which are toothless. The complexity of teeth may also be partially attributable to differences in amelogenin between different groups. It seems that the rate of accumulation of non-synonymous mutations is inversely related to the complexity of tooth forms in an individual (Mathur and Polly 2000);

though the authors of this study suggest that the trend may be largely dependent on a high % protein divergence in monotremes, which have vestigial adult dentition. Additionally, in the species with more complex teeth, amelogenin can be alternatively spliced to give differing protein products. Multiple transcript isoforms have been detected in mice (7), rats (5) and pigs (6). In pigs, the different transcript lengths have been associated with specific enamel layers within a single tooth, supporting the theory that the transcripts work together to contribute to tooth complexity (Mathur and Polly 2000).

The importance of the differences in dentition between taxonomic groups is clear even without such molecular studies. As one of the most durable biological structures, teeth form a significant portion of the vertebrate fossil record. Individual teeth have tracked both the diversification and dispersion of ancient mammal species; and the therian mammals (marsupials plus placentals) are defined by the tribosphenic molar (Benton 2000). More recently, differences in tooth thickness, which may be a direct correlate of enamel deposition, between apes and ancient and modern man, have fuelled speculations on the behaviour and ecology of early man (Hillson 1996). The layering of different enamel forms in humans is more akin to gibbons and *Sivapithecus* (potentially a common ancestor of higher apes) than to modern day gorillas and chimpanzees with orangutan falling somewhere in between (Hillson 1996). Whilst some have explained this in terms of adaptation to specific diets or selection on body size (reviewed in Hillson 1996), others have hypothesised a relaxation of selection on human dentition, and linked this to the invention

of cooking (Brace 1963). Within the genus Homo, there is a downward gradient in cheek tooth diameter from H. erectus through Neanderthal man to modern humans (Hillson 1996).

IV.1.5 Phylogenetic relationship of amelogenin sequences

Figures IV.2A and IV.2B are neighbour joining trees based on non-synonymous and synonymous changes in the amelogenin coding region, respectively (Modified Nei-Gojobori method (1986), MEGA2 (Kumar et al. 2001)). The sequences are from Iwase et al. (2003). The distances on Figure IV.2 are scaled by the total number of sites, not the number of potentially synonymous and non-synonymous sites (AMEL ratio 1: 2.3). In a neutrally evolving gene, one would expect the branches on the non-synonymous tree to be roughly twice as long as on the synonymous tree. For the majority of lineages it is clear that the number of non-synonymous changes is around half the number of synonymous changes. This is good evidence that the sequences have remained under functional constraint and that most amino acid changing mutations have been censored by selection before they could become substitutions. However, the single branch leading to the human and chimpanzee Y sequences appears somewhat anomalous. Instead of the proportion of non-synonymous changes being a fraction of the synonymous changes, amino-acid replacements seem over-represented in this lineage. Lahn and Page (1999) calculated AMELX/Y pairwise Ka/Ks as 1.0 (0.07/0.07), which, along with the preponderance of pseudogenes in stratum 4, has been

taken as strong evidence that AMELY is no longer under functional constraint.

The high proportion of genes in stratum 4 that have become pseudogenes since their restriction to the Y chromosome and the observed degeneration of Y linked genes on older strata, would predict the demise of the Y linked copy of amelogenin. Other genes retained on both X and Y (e.g. SMCY, UTY, USP9Y) are ubiquitously expressed. However, this gene is still transcribed from the Y and its complete absence produces a phenotype that would seem to be disadvantageous. The Ka/Ks ratio is around one and, whilst there is no specific theoretical criterion detailing when Ka/Ks 'around one' becomes 'above one', this may yet represent the neutral evolution of a largely unimportant gene or may mask the signal of natural selection operating in the human lineage.

The aim of this study was to investigate the molecular evolution of sex linked amelogenins in a range of primates to describe the difference between X and Y, the degree to which other primate AMELY genes had altered and, if possible, test for presence of positive selection.

What was discovered was somewhat different.

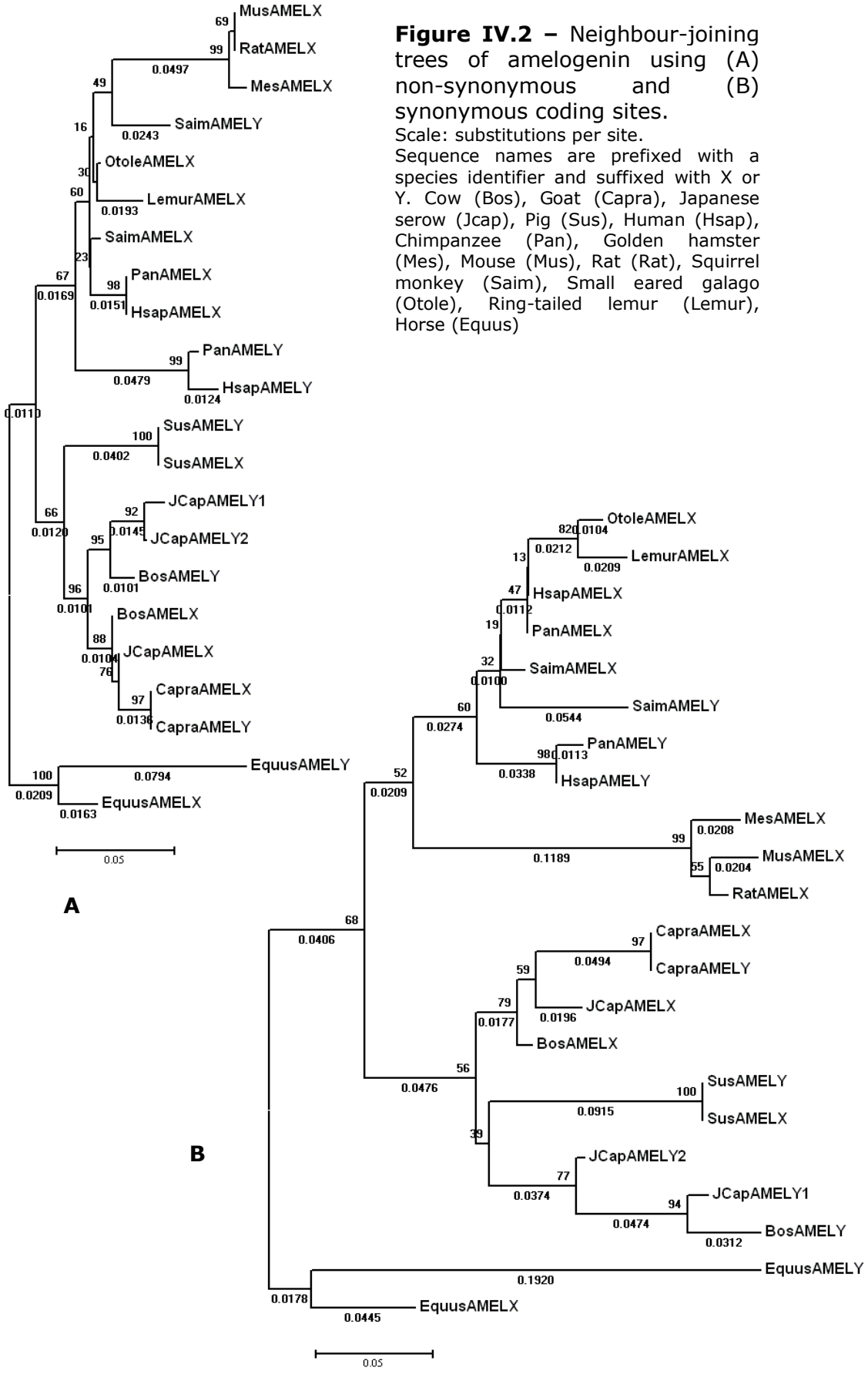


Figure IV.2 – Neighbour-joining trees of amelogenin using (A) non-synonymous and (B) synonymous coding sites.
Scale: substitutions per site.
Sequence names are prefixed with a species identifier and suffixed with X or Y. Cow (Bos), Goat (Capra), Japanese serow (Jcap), Pig (Sus), Human (Hsap), Chimpanzee (Pan), Golden hamster (Mes), Mouse (Mus), Rat (Rat), Squirrel monkey (Saim), Small eared galago (Otole), Ring-tailed lemur (Lemur), Horse (Equus)

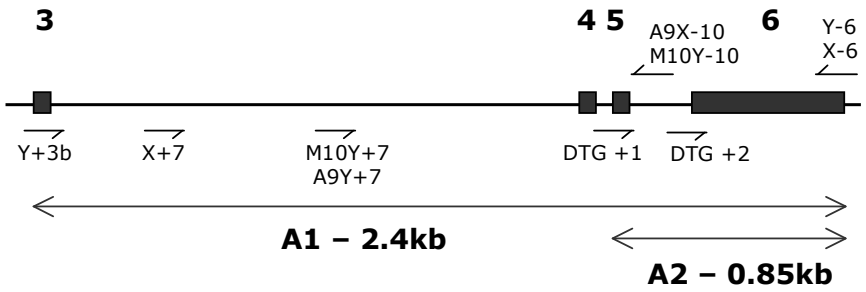
IV.2 Materials and Methods

New partial sequences from X and Y linked amelogenin were obtained from a range of primates to fill out the primate Y phylogeny of these genes. Due to indels in other sequences, only the horse, mouse and primate sequences in Figure IV.2, which included non-coding as well as coding sequence, were retained to be aligned with new sequences.

IV.2.1 PCR, cloning & sequencing of amelogenin

Partial amelogenin sequences corresponding to 56bp of exon 5, the entirety of intron 5 and 481bp of the large exon 6 were amplified in a range of primates. The primers used are listed in Table IV.1 and their priming sites are shown on Figure IV.3. Primer DTG+1 was designed on human Y sequence to be Y specific as it had 3 base differences from the human X sequence. PCR and cloning was carried out as in Chapter III. All

Figure IV.3 – Positions and directions of primers used around exons 3-6. The positions of the final alignment are marked below



successful PCR products that resembled amelogenin sequence were cloned using the TOPO cloning Kit (Invitrogen). Additional 5' sequencing from exon 3 to exon 6 was carried out for several clones that proved difficult to

position phylogenetically and this is described in Table IV.2 in the results section.

Table IV.1 – PCR & sequencing primers used.
* (Iwase et al. 2003) †Used in PCR

Primer	Sequence	Annealing temperature
Y+3b†	CCTCATCCTGGGCACCCTGGTTATATCA*	66.1ºC
DTG+1†	ATCTTTCTCTTAAGGTGCTCACCC	54.1ºC
Y-6†	CACTTCCTCCTGCTTGGTCTTGTC*	58.0ºC
X-6†	CACTTCCTCCCGCTTGGTCTTGTC*	61.9ºC
X+7	AGAGACTCTGCAGAACAATGGAATG	55.6ºC
M10Y+7	GAATCCCCAAACCTAAGGTTAAC	52.1ºC
A9Y+7	GAACCCTCAAACCTGAGGTTAAC	52.8ºC
M10Y-10	TTAGTGTCTGTATGTGGAGTACGC	50.7ºC
A9X-10	GATTAGTGTCAGTATGTGGAGTAGAC	48.0ºC
DTG+2	TGGCCATAATGGCAAAGACAACAC	59.6ºC

IV.2.2 Assembly and analysis of alignments

An alignment of all sequences from X and Y was assembled in a single Gap4 database (Staden, Beal, and Bonfield 2000). Using ProSeq (Filatov 2002), this alignment was then amalgamated with published sequence (Iwase et al. 2003) (used in Figure IV.2), which had been downloaded from Genbank to form a 'long' alignment. Several sub-alignments were constructed. The first (A1) featured the published sequence and the sequence from clones that had received additional sequencing (see Table IV.2). The second alignment (A2) included both non-coding and coding portions shared by all sequences and the third alignment (A3) contained just the coding region shared by all sequences.

Due to the high homology of X and Y coding sequences but low homology of non-coding sequences, the PCR primers used often amplified two products of similar lengths. Where sequences of these PCR products could not be confidently matched to either human X or Y sequence, additional sequencing covering the region from exon 3 to exon 5 was conducted. The cloning of all sequences insured that each contig was from a single chromosome.

Each alignment was checked by eye in ProSeq. Using alignment A1, the number and length of insertions and deletions in coding and non-coding regions were scored. For alignment A2, the range of base compositions of coding and non-coding regions across the sequences was measured.

IV.2.3 Building a Phylogeny

For the overall alignment and each sub alignment, phylogenetic trees were constructed using the neighbour joining (NJ) and maximum parsimony (MP) methods implemented in MEGA2 (Kumar et al. 2001). Bootstrap support for the trees came from 1000 replicates for the NJ tree and, when the number of sequences was not too great, the MP tree also. For alignment A3, a consensus of the equally most parsimonious trees was taken.

IV.2.4 Tests for recombination

Because of the historical association of the amelogenin locus around an ancient PAR boundary, the perceived risk of gene conversion between X and Y chromosomes within a species and the apparent phylogenetic relationship of coding sequences, the entire data set were tested for the signal of recombination. Initially, the four-gamete test was used on the entire set of sequences (Hudson and Kaplan 1985) using DNAsp (Rozas and Rozas 1999). This test compares a pair of polymorphic sites (or entire loci at a larger scale), if each site has two alleles (e.g. (a,t) & (t,c)), then recombination may create four possible haplotypes (at, ac, tt & tc). Without recombination, each allele will only exist in *cis* with alleles of the chromosome on which it arose. This test assumes that there are an infinite number of sites i.e. each site only mutates once, and it may perform badly in regions of repetitive or low complexity sequence, which may be poorly aligned. Subsequently X-specific, Y-specific and outgroup sub-sets of sequences were also tested.

As the results of the four-gamete tests were positive for the entire data set and most sub-sets of sequences (see results), the programme TOPALi (Milne et al. 2004) was used to analyse the support for recombination in alignment A1, using the following three methods.

Under the Difference in sums of squares method (Dss) (McGuire and Wright 2000), a window of part of the alignment is divided into two halves. For the first half, a best tree is generated by minimising the

overall branch lengths according to the un-weighted least squares criterion, which is the sum of squared differences between pairwise comparisons and distances on the tree. Then, fitting the second half of the alignment window to the same tree generates a second sum of squares. The difference in sums of squares between the two windows is the Dss statistic and should be greater across areas featuring recombination. The significance threshold of the statistic is generated in TOPALi by parametric bootstrapping. As the 600bp exon 6 was suspected to be recombinant relative to the preceding introns in some sequences, the programme was run using a window size of 800bp and the default step size of 10bp.

The Probabilistic Determination Method (PDM)(Husmeier and Wright 2001) also uses a sliding window approach but generates trees independently for each half of the window by sampling from a posterior distribution using Markov Chain Monte Carlo simulation (Husmeier and Wright 2001). This distribution is expected to change across a recombinant region and a local measure of this change is plotted against the sequence length to present a graphical prediction of recombination points. Again, the programme was run using a window size of 800bp and a step size of 10bp. Due to the computer time required to run this analysis, only the four anomalous sequences were used as input.

The Hidden Markov Model (HMM) (Husmeier and McGuire 2003) method of recombination detection, which can only be used on four sequences, estimates the support along the alignment for each of the three possible unrooted, bifurcating tree topologies. "Statistical significance is assessed

by (posterior) probabilities assigned to each topology for each position in the alignment." (TOPALi user manual). This method was used on the two anomalous 'Y' sequences from Ateles (A09) and Macaca (M10) and their con-specific 'X' homologues, which were suspected of recombination. As a comparison, four other divergent sequences from the X chromosomes of mouse, galago (Lemur family) and human and from the Y chromosome of human were tested without an *a priori* reason to suspect recombination.

IV.2.5 Analysis of parsimoniously informative sites

An analysis of the parsimoniously informative sites in alignment A2 (840bp) was conducted to further characterise the evolution of the amelogenin genes. The purpose of the analysis was to try and differentiate between recombination and homoplasy. Using the horse, mouse and lemur sequences as outgroups, sites were scored on their inclusion of key sequences in an X-derived, Y-derived or New World-derived state. For example, a site was scored as Y derived if all Y sequences, including the key sequence, shared a nucleotide which was different from that shared by all the Xs and the ancestor. Sites fitting none of these patterns were classified as mixed. Where there were differences in the outgroups, the ancestral (at the time of the split of lemurs from anthropoid apes) sequence was decided according to the following hierarchy: (i) the mouse and at least one lemur sequence were the same; (ii) one lemur and one horse sequence were the same and other ancestral sequences did not match; (iii) if lemurs were un-informative, mouse took precedence over horse.

IV.2.6 Mutation analysis

The programme baseml (PAML, (Yang 1997)) was used to conduct maximum likelihood reconstruction of ancestral sequences at each of the nodes in the best tree for alignment A1. The output of this programme assigned mutations to specific lineages. From this information, the frequencies of each type of mutation along each branch were scored.

IV.2.7 Maximum Likelihood testing of models of evolution

To assess the relative importance of constraint in X and Y lineages and to test for the presence of positive selection operating, the likelihoods of several evolutionary models, testing either variation amongst lineage or variation among sites, were generated and compared. The programme codeml, part of the PAML package (Yang 1997) was used to calculate the likelihoods of the models on the coding sequence forming alignment A3.

Two forms of tests were made. The first estimated different rates of Ka/Ks, or ω, for different lineages in the phylogenetic tree relating the sequences. The second approach estimated different ω values at different codon sites within the gene, averaging across the tree. Both are discussed in Chapter III, along with references to studies suggesting that the tests for both features at once are unconservative. The test used was the Likelihood Ratio Test which, given the nested nature of the models, follows a χ^2 distribution under the null hypothesis that the more parameter rich model explains the data no better.

Firstly, a model allowing the ω ratio along the monkey and ape Y lineage to differ from the rest of the tree was compared with a model in which just one ω ratio was estimated for the entire tree. Secondly, heterogeneity in ω was tested for amongst the codon sites within the Y lineage. Model M3, allowing two classes of ω, was tested against Model M0, which allowed just one ω ratio for all sites. The test for positive selection used models M14 and M15 as in Chapter III. A significantly better fit to the data of Model M15, which allows a separate class of ω>1.0, over M14, is an indicator that a proportion of sites are accumulating non-synonymous substitutions at greater than the neutral rate.

IV.3 Results

IV.3.1 Sequencing

The new AMEL sequences are listed in Table IV.2. The codes correspond to particular DNA samples and are used to identify sequences in the phylogenetic trees below. Using primers DTG+1/Y-6 (X-6 for Pongo), no AMELY homologue was obtained for Presbytis or Leontopithecus; however, both species produced AMELX homologues. For Colobus, Gorilla, Ateles and Macaca, multiple products were evident in a single PCR (as double peaks in direct sequencing) and these products were separated as part of the cloning of all products. Note that the original Macaca (M10) and Ateles (A09) products using forward primer DTG+1 were difficult to assign as X or Y homologues based on pairwise similarity of phylogenetic position. Hence, longer PCR products (which had not been obtained from other species) were sequenced. The primers X+7, M10Y+7, M10Y-10, A9Y+7 and A9X-10 were designed to complete this sequencing.

IV.3.2 Alignments and phylogenies

The extended sequences from Macaca (M10_APb, M10_AOa)[†] and Ateles (A09_AMc, A09_ANb) were added to the published sequence used in Figure IV.2. Retaining primate, horse and mouse sequences, the

[†] APb, AOa, AMc & ANb are the names of specific cloned PCR products and are used here because at this stage, the homology of these sequences was ambiguous

alignment length was 2415bp and is referred to as A1. Figure IV.4 shows the neighbour joining tree (Tamura Nei (1993) distance) of alignment A1.

Table IV.2 – Sequencing of AMELX/Y
* Sequences which initially proved difficult to associate with human X or Y and received additional sequencing

Code	Species	Common name	Primers used	Product length (bp) X	Y
C71	*Colobus sp.*	Black & white colobus monkey	DTG+1 /Y-6	731	733
P70	*Presbytis entellus*	Hanuman langur	DTG+1 /Y-6	749	-
G73	*Gorilla gorilla*	Gorilla	DTG+1 /Y-6	749	751
P11	*Pongo pygmaeus*	Orangutan	DTG+1 /X-6	-	752
G72	*Hylobates lar*	Lar gibbon	DTG+1 /Y-6	-	746
T65	*Leontopithecus rosalia*	Golden lion tamarin	DTG+1 /Y-6	750	-
A09	*Ateles geoffreyi*	Black spider monkey	Y+3b /Y-6	2166*	1765*
M10	*Macaca fasicularis*	Macaque	Y+3b /Y-6	2186*	1985*

All nodes have 95% bootstrap support or greater and the topology agrees perfectly with the relationships of the primates (Purvis 1995) following the split of X and Y after the divergence of the lemurs from the Anthropoid primates. This tree clearly places one sequence each from Macaca (M10_AOa) and Ateles (A09_ANb) in the X lineage and one from each in the Y lineage (M10_APb, A09_AMc). In both lineages, the Ateles (A09) sequence forms a sister clade with the other New World (NW) monkey, Saimiri whilst, the sequences from Macaca (M10), an Old World (OW)

monkey, fall as expected between the NW monkeys and the apes. The maximum parsimony tree (not shown) shared the same topology.

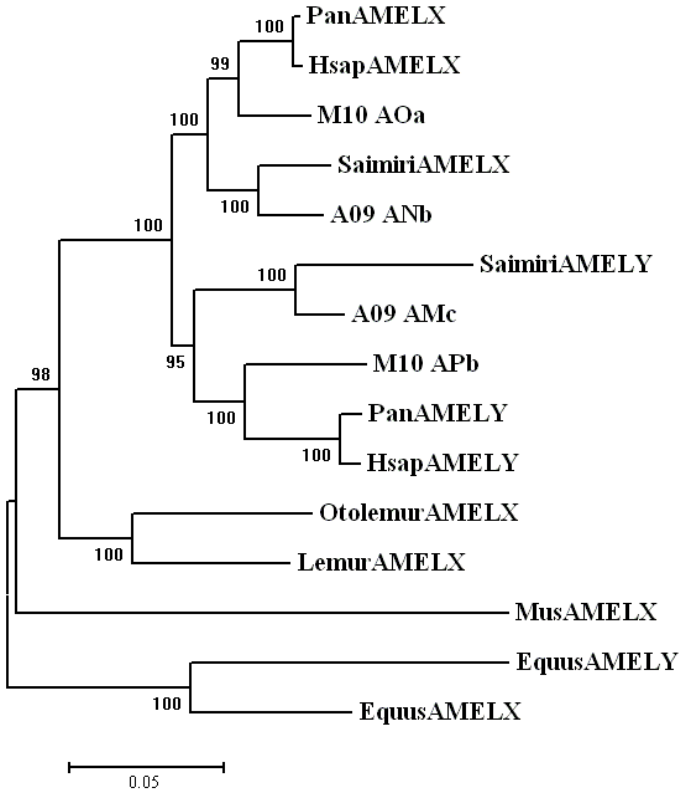


Figure IV.4 - Neighbour joining tree of alignment A1. Scale: substitutions per site

A second alignment (A2) was created by adding the sequences from the other species (see appendix). The length of this alignment was 840bp and, starting in exon 5, spanned intron 5 (291bp) and the larger part of exon 6. Phylogenetic trees were created from this alignment using both the neighbour joining and maximum parsimony methods. Both trees bore some resemblance to that in Figure IV.4 but also contained a number of distinct incongruencies. Figure IV.5 shows the bootstrap consensus of the neighbour joining tree. Whilst the OW monkey and Ape X and Y sequences have split into two distinct clades, all the NW monkey sequences have

been excluded and the node of (OW X&Y, NW X, NW Y) is unresolved (bootstrap support less than 50%). Another unresolved node is that of the human, chimp and gorilla Y sequences. Several other nodes are misplaced according to the accepted relationship of these species. In the Y lineage, Colobus (C71), an OW monkey, forms an exclusive clade with Hylobates, the gibbon (an ape). In the X lineage, Macaca and Presbytis form a clade to the exclusion of Colobus. The phylogeny of Purvis (1995) would place Colobus and Presbytis as nearest neighbours. It should be noted that the bootstrap support for this last node is only 60% and for the node excluding the NW monkey sequences from the OW X and Y lineages is 61%. Both may be fallacious.

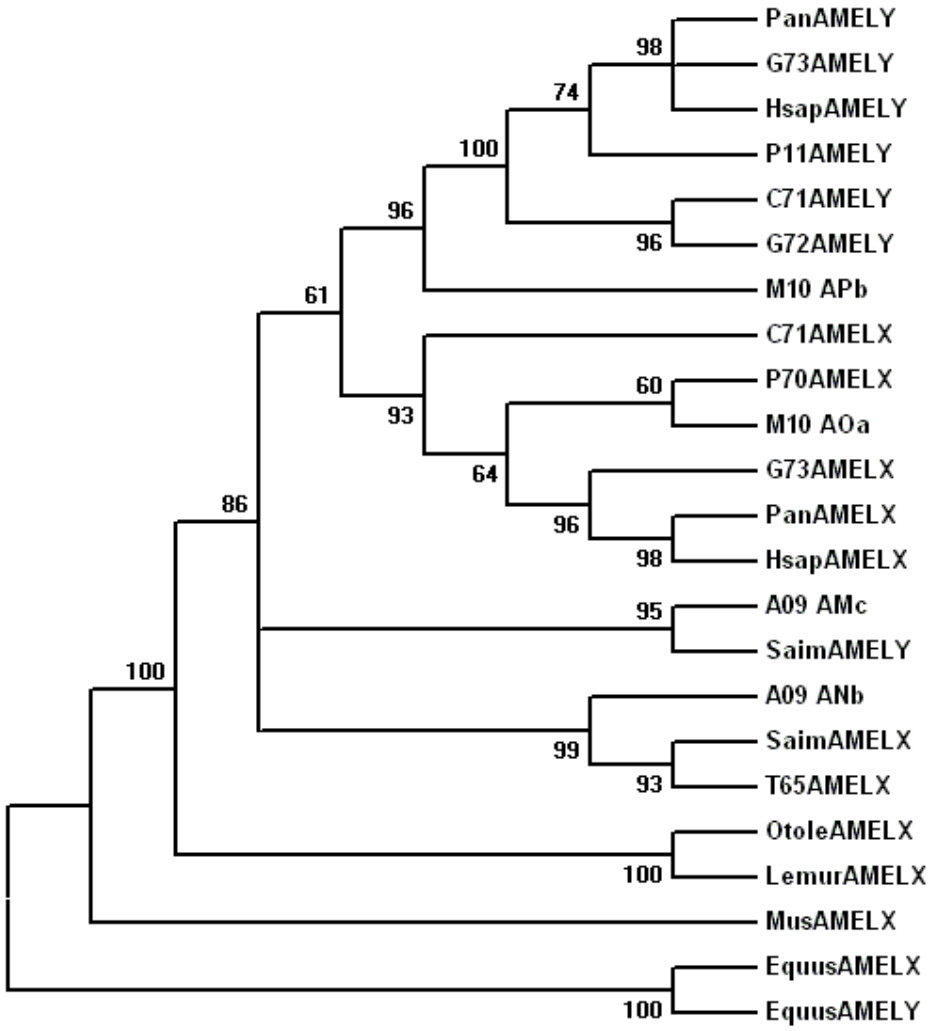


Figure IV.5 – Bootstrap consensus tree of alignment A2 (NJ method).

Doubt in the phylogeny obtained from alignment A2 is increased by the tree in Figure IV.6. This tree is a consensus of the equally most parsimonious trees for alignment A2. Whilst the Y clade has been recovered as per the longer alignment, A1, the X clade has been broken, with the NW X cast adrift and clustering immediately outside of the Y group. Although, this is with marginal bootstrap support of 50%. Again,

the positions of the Hylobates and Colobus Y sequences are non-conforming, as are the OW monkey X sequences within the X lineage.

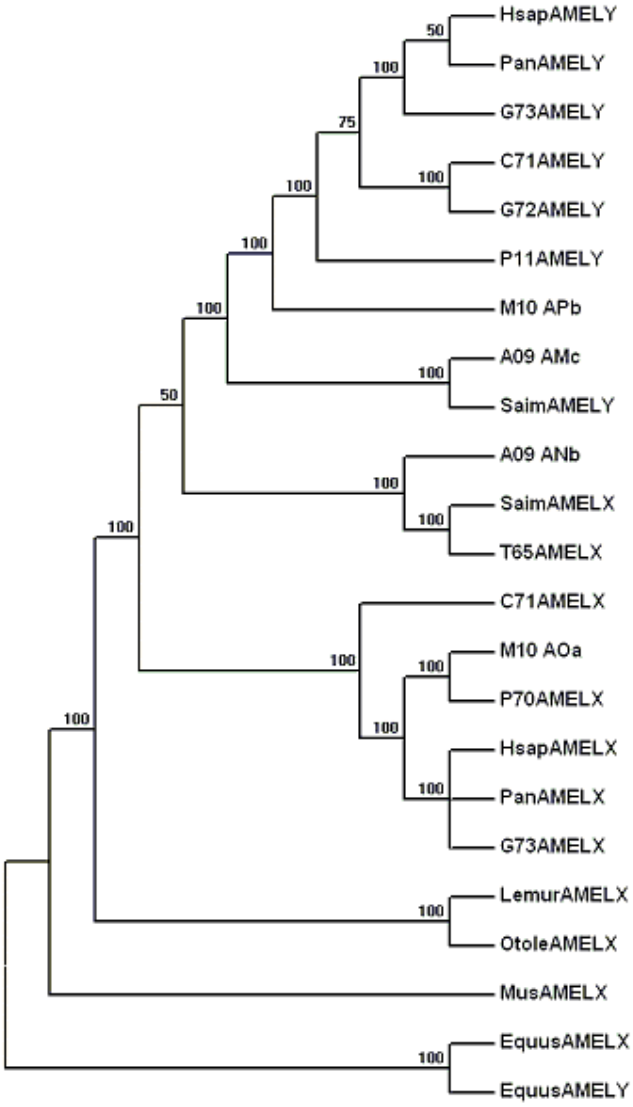


Figure IV.6 – Consensus tree of sequences in alignment A2 (Parsimony method) Because of the number of sequences, branch labels are not bootstrap values but denote the proportion of equally parsimonious trees sharing that node. Node labels are as in Figure IV.2, with new sequence labelled according to Table IV.2.

The next phylogenetic tree, Figure IV.7, was generated by the NJ method from the shortest alignment, A3, comprising only the coding sequence of exons 5 and 6, and with a length of 537bp. This tree features the collapse of many internal nodes, due to low bootstrap support and in other places formation of nodes not featured in other trees e.g. [(HsapAMELY, G73AMELY), PanAMELY] and [(M10_APb, C71AMELX), all else].

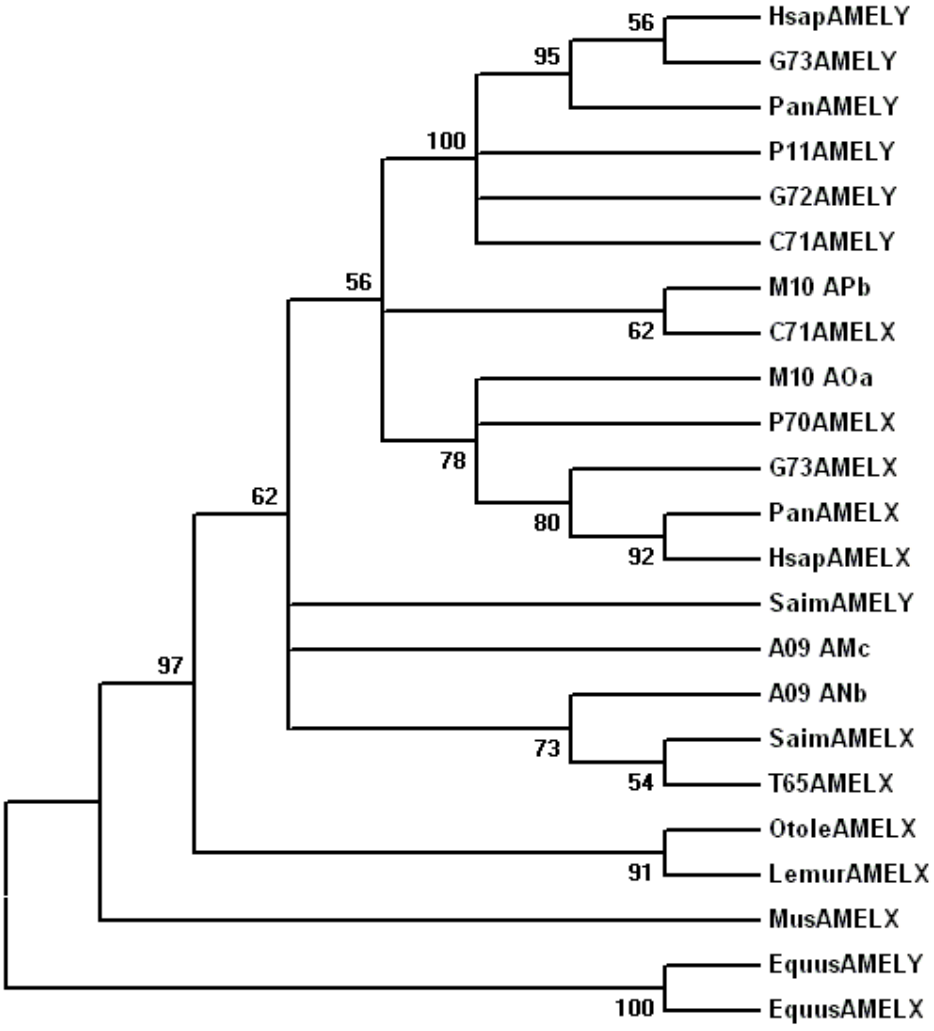


Figure IV.7 – Bootstrap consensus tree of alignment A3 (NJ method)

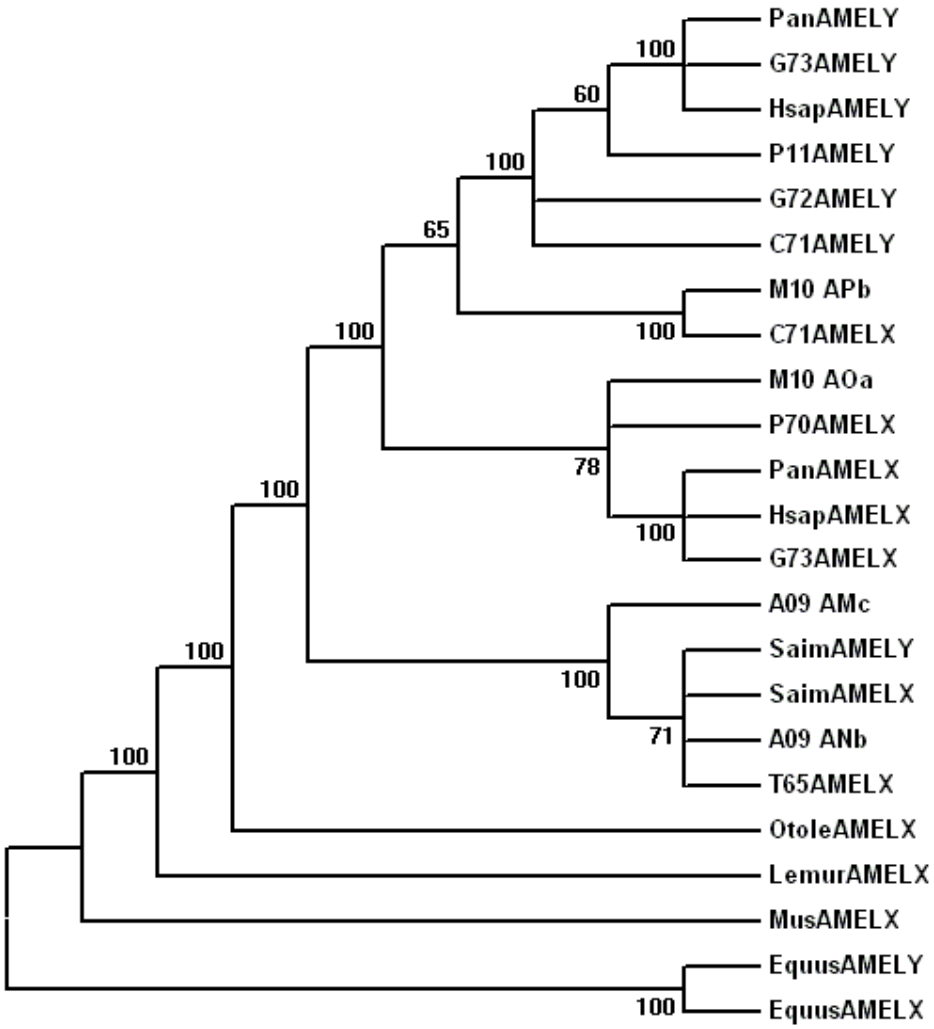


Figure IV.8 – Consensus tree of alignment A3 amino acid sequence (Parsimony method) In this tree, node labels are not the bootstrap consensus but a consensus taken across all equally parsimonious sites.

Finally, Figure IV.8 illustrates the tree generated by using the parsimony method on amino acids of the coding alignment A3. Whereas again there are general clades for the Old World apes and monkey X and Y sequences, the sequences A09 AMc and M10 APb, which were strongly Y clustered in Figure IV.4, have altered positions. The Macaca (M10) 'Y' sequence forms a clade with Colobus (C71) X and the Ateles (A09) 'Y' sequence along with the Saimiri Y sequence are now in a clade comprising both the X and Y

sequences from all the New World species (Ateles, Saimiri, Leontopithecus (T65)).

Whilst incongruencies in a tree using less sequence data (alignment A3 compared to A1) may be due to a lack of phylogenetic signal, the consistent movement of certain Y sequences (namely A09 AMc from Ateles and M10 APb from Macaca) to positions in the tree closer to the X sequence from the same species, raises the possibility of the occurrence of gene conversion between X and Y copies.

IV.3.3 The signal of recombination (gene conversion)?

The four-gamete test for recombination was used on the full set of sequences and the results were surprising (Table IV.3). With individual X and Y sequences from divergent species, no recombination was expected and yet the four-gamete test detected a very strong signal. To try to localise where in the dataset the signal was coming from, several sup-groups were also tested. Again surprising was that most sub-sets gave a signal of recombination between sequences, including those from Y only sequences. The Macaca and Ateles sequences were removed from the X and Y sub-sets and for the X group, but not the Y, this removed the signal of recombination (though this does not rule out the remaining X sequences as recombinants). Interestingly, using the small number of both X and Y sequences from all the NW monkeys or all four Ateles and Macaca sequences, still gave a signal of recombination.

Table IV. 3 – Results of the four-gamete test on alignment A2
NW – New World monkeys from genera Ateles, Saimiri & Leontopithecus

Sequences	Number of sequences	Minimum # of recombination events
Full data set	23	33
Outgroups (mouse, horse, lemur)	5	12
Monkey & ape X	9	2
Monkey & ape X (less NW & Macaca)	5	0
Monkey & ape Y	9	13
Monkey & ape Y (less NW & Macaca)	6	4
Ateles & Macaca (from X & Y)	4	1
All NW Monkeys (from X & Y)	5	6

To better characterise the position of any recombination, the programme TOPALi (Milne et al. 2004) was used to search for recombination along alignment A1 with the three different methods explained in the methods section. Figure IV.9 illustrates the results of the PDM analysis on alignment A1. Under this analysis only four sequences are used. The line graph in the background of Figure IV.9 represents the signal of recombination along the sequence with the dashed line representing a posterior threshold value which, when surpassed, would indicate recombination. In Figure IV.9 this does occur around 1800 bases into the alignment. The two trees in the foreground of Figure IV.9 illustrate the programme's prediction for the relationships of the sequences based on the signal and position of recombination. Whereas the first 1800bp fit the topology of the entire alignment shown in Figure IV.9, the PDM analysis suggests that for the last 600bp of the alignment, mostly comprised of

exon 6, the relationship of the Macaca and Ateles sequences is different with the 'X' and 'Y' being more closely related within a species than are the homologous chromosomal sequences between species.

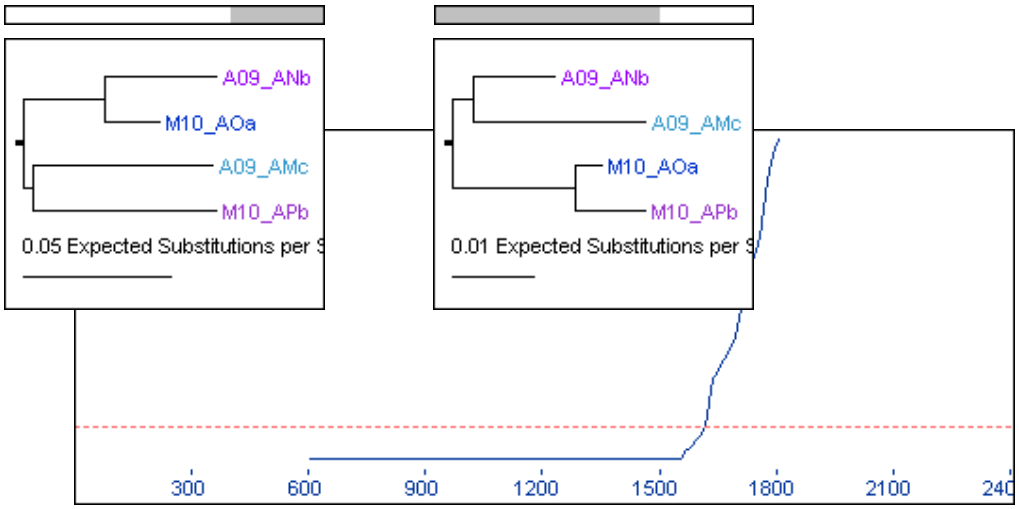


Figure IV.9 – Output from TOPALi of PDM run on the four anomalous sequences from Ateles (A09) and Macaca (M10) The two trees shown are the best supported tree either side of the recombinational breakpoint.

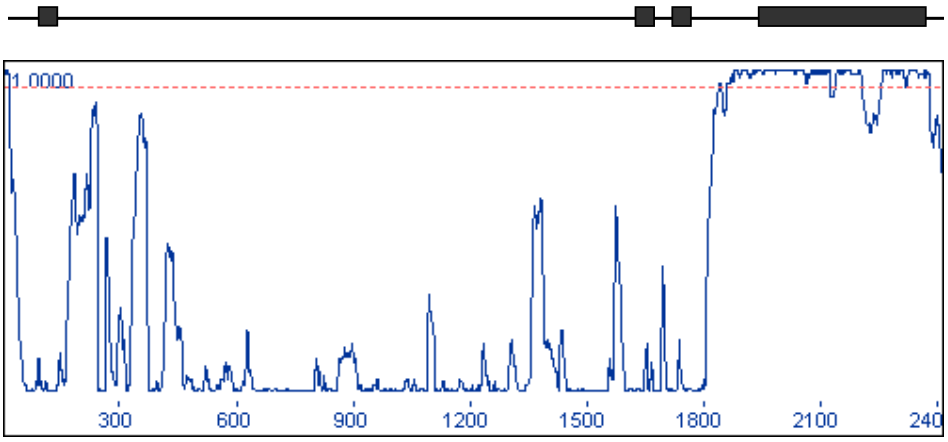
Under the Dss method on all sequences in alignment A1, the Dss statistic exceeded the recombination detection threshold in four separate places along the alignment, most significantly around 1800bp in (results not shown).

Under the HMM method, support was measured along the alignment for three different topologies of four sequences. This method was applied to the four anomalous sequences from Macaca and Ateles and also to four other sequences, which had been placed as expected in most of the phylogenetic trees (HsapAMELX, HsapAMELY, MusAMELX & OtoleAMELX). Figures IV.10 and IV.11 show the HMM output from TOPALi for both of these runs. The graphs show the support for each of the three possible

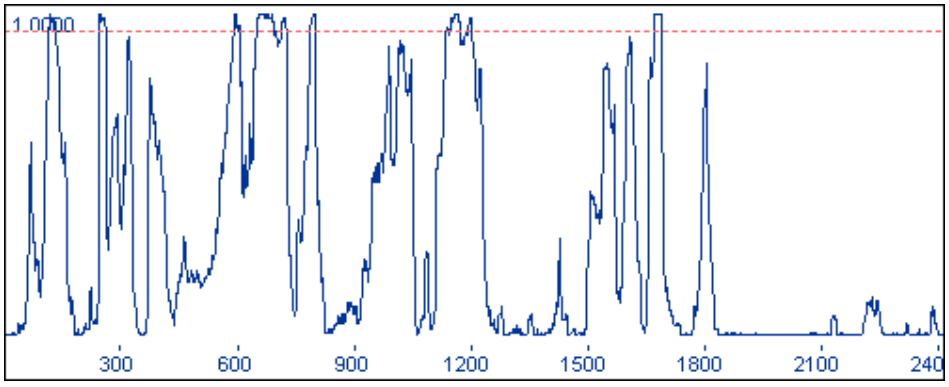
tree topologies along the length of the alignment. For the Macaca and Ateles sequences (Figure IV.10), Topologies 2 ([(A09X, M10X), (A09Y, M10Y)] as per Figure IV.4) and 3 [(A09X, M10Y), (M10X, A09Y)] seem to be evenly supported up to 1800bp into the alignment. After this point, support switches almost exclusively to Topology 1 [(A09X, A09Y), (M10X, M10Y)], grouping the within species sequences with each other as per the PDM and Dss analyses. However, the HMM results for four other sequences are surprisingly similar. Figure IV.11 shows the output from TOPALi of the same analysis run on the human Y sequence and the human, mouse and galago (Lemur family) X sequences. Again there is a strong change in support for topologies around 1800bp into the alignment. Before this point, the accepted relationship (as per Figure IV.4) is supported, but after this point, support changes, to a topology grouping the human and mouse X sequences to the exclusion of the human Y sequence.

IV.3.4 Analysis of indels in alignment A1

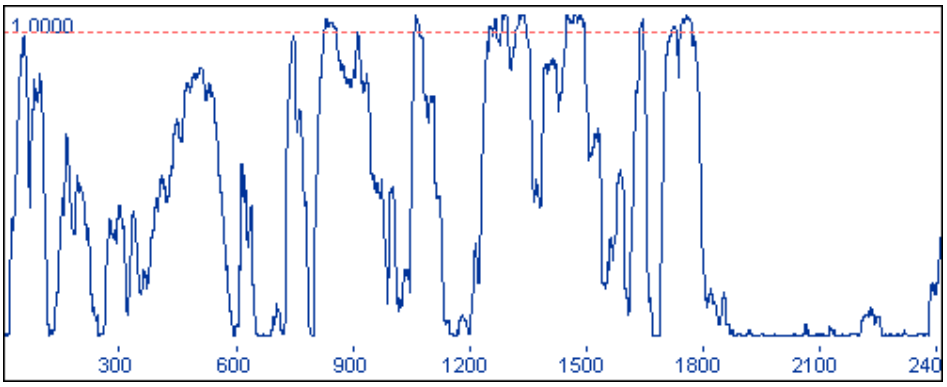
In the 2kb alignment A1, there were 126 indels but only 10 (8%) occur within coding regions (24% of the alignment). Of the coding indels, four of the five frameshift indels occur in the 45bp exon 4: EquusY (–2bp, -4bp); EquusX (+4bp); NW Y lineage (+1bp). As this exon is alternatively spliced from the majority of human AMELX transcripts (Salido et al. 1992), indels here may not affect the coding sequence downstream. The other frameshift was a 1bp insertion at the end of the long exon 6 in Equus AMELY. The other five exonic indels are all in-frame.



Topology 1: (A09_ANb, A09_AMc), (M10_AOa, M10_APb)

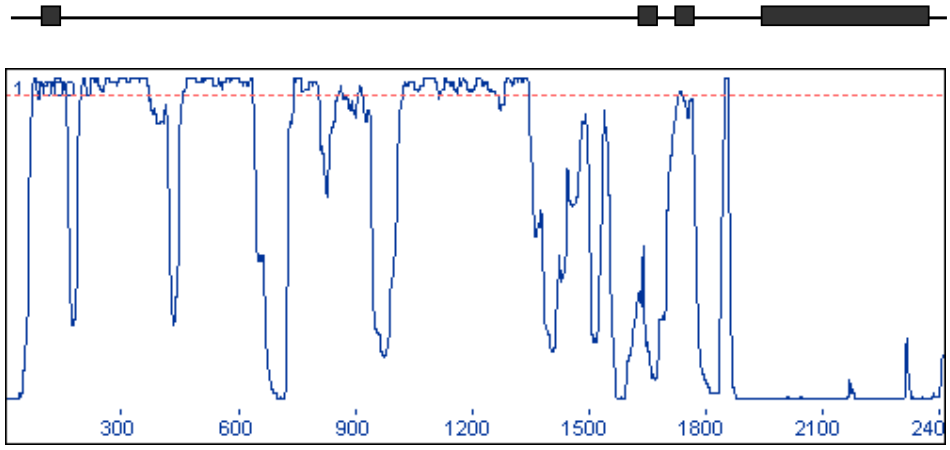


Topology 2: (A09_ANb, M10_AOa), (A09_AMc, M10_APb)

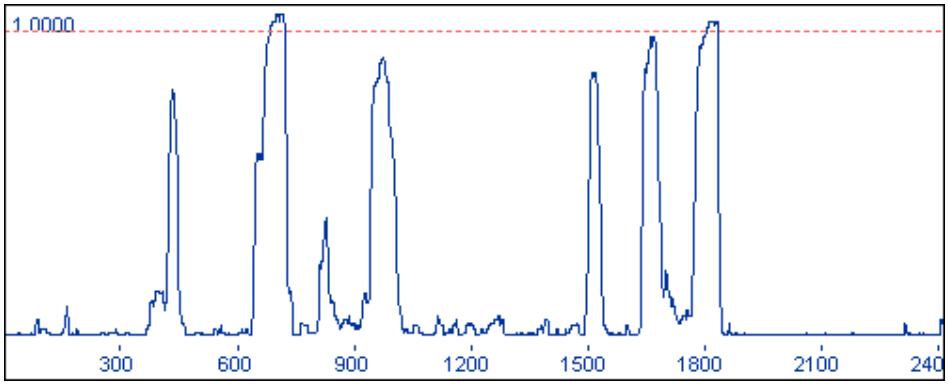


Topology 3: (A09_ANb, M10_APb), (A09_AMc, M10_AOa)

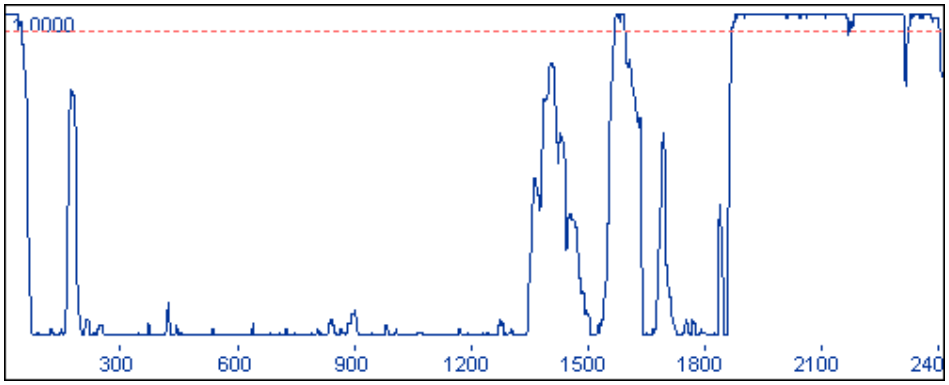
Figure IV.10 - Output from TOPALi of HMM run on the four anomalous sequences from Ateles (A09) and Macaca (M10)
Exon structure (3-6) given above.



Topology 1: (HsapAMELY, HsapAMELX), (MusAMELX, OtolemurAMELX)



Topology 2: (HsapAMELY, MusAMELX), (HsapAMELX, OtolemurAMELX)



Topology 3: (HsapAMELY, OtolemurAMELX), (HsapAMELX, MusAMELX)

Figure IV.11 - Output from TOPALi of HMM run on Human X & Y, Mouse X and Otolemur X. Exon structure (3-6) given above.

The reduction of frameshift indels in the exons is a strong indicator that most of these sequences either still are or have been until very recently, under purifying selection.

IV.3.5 Base composition and codon usage in alignment A2

There is a strong difference in the base composition of exons 5 and 6 compared to intron 5 of this gene. Table IV.4 illustrates the massive % GC bias in the exons (all sites) relative to the introns. The bias is common to all sequences in all species and, judging by the overlapping frequencies of exonic A, G and T, is due to a high proportion of C nucleotides. The range of both overall exonic GC% and that at third positions of codons (GC3%) is at the very high end of the distribution in the human genome (Eyre-Walker and Hurst 2001). The range of GC3% in primate AMELX sequences is 90-93. For the Y lineage, it is 83-91. No X sequence has a lower GC3% than its Y homologue (values non-independent). Horse GC3% values are 90% and 79% for X and Y respectively. Conversely, the intronic GC% content seen here (30-36%) is in the lower third of values for blocks of contiguous sequence around the genome (Duret, Semon et al. 2002).

Table IV.4 – Range of base composition in alignment A2 (all sites)

	bp	%A	%T	%G	%C	%GC
Non-coding	303	30-36	30-35	13-18	14-20	30-36%
Coding	537	19-23	15-19	18-20	40-44	57-64%

The base compositional bias of the amelogenin exons 5 and 6 relative to intron 5 may be largely explained by their unusual amino acid content.

The amelogenin protein includes an exceptionally high proportion of proline residues, an amino acid usually found at much lower frequency. Table IV.5 lists the numbers of Proline, Glutamine, Leucine, Histidine, Valine and Isoleucine residues found in each of the sequences making up alignment A3. Also given are the numbers of each synonymous codon used. Methionine, which was the fifth most common amino acid but only uses one codon, was omitted from the table. The proportions of Pro, Gln and His are all well above the genomic human frequencies (taken from http://www.kazusa.or.jp/codon/ and based on Genbank release 144.0 (12th November 2004)). Some of the Pro (P) and Gln (Q) residues in the amelogenin X and Y sequence make up 15 and 13 QP pairs, respectively, in the latter half of the peptide sequences shown below.

Human AMELX peptide sequence (Genbank: NP_001133)

```
  1 mgtwilfacl lgaafamplp phpghpgyin fsyevltplk wyqsirppyp sygyepmggw
 61 lhhqiipvls qqhppthtlq phhhipvvpa qqpvipqqpm mpvpgqhsmt piqhhqpnlp
121 ppaqqpyqpq pvqpqphqpm qpqppvhpmq plppqpplpp mfpmqplppm lpdltleawp
181 stdktkreev d
```

Human AMELY peptide sequence (Genbank: NP_001134)

```
  1 mgtwilfacl vgaafamplp phpghpgyin fsyevltplk wyqsmirppy ssygyepmgg
 61 wlhhqiipvv sqqhplthtl qshhhipvvp aqqprvrqqa lmpvpgqqsm tptqhhqpnl
121 plpaqqpfqp qpvqpqphqp mqpqppvqpm qpllpqpplp pmfplrplpp ilpdlhleaw
181 patdktkqee vd
```

Together these values suggest that, whilst the amino acids used by amelogenin inflate the GC%, historically, the silent sites (GC3%) also had high GC% but this has declined in the Y lineage. This is consistent with the findings of Eyre-Walker (1993), comparing the ratio of GC to AT mutations from several X and Y linked genes. However, why the GC%

content of third codon positions in amelogenin is not more similar to the intron content is not clear.

Table IV.5 also suggests a further potential link between the amino acid content of the gene and the mutation pressures affecting it. For each of the amino acids listed, there is a strong preference for some synonymous codons over others. In most cases, the preference is similar to that for the genome as a whole e.g. the nucleotide triplet CAA occurs at 26% of Gln sites in the genome and 22% of Gln sites in AMEL on average. However, there seems to be an additional pattern in this table. The least common codon of Proline, CCG, is used more often in the X lineage and the outgroup lineages than it is in the Y lineages. This difference is not made up by an increase in the use of a synonymous codon and, judging by the reduction in the total number of Prolines, of which there are less in the Y lineages, reflects a series of non-synonymous mutations of this codon to another. While all the numbers in Table IV.5 are phylogenetically related and therefore non-independent, this difference in Proline codon frequency does seem to increase in strength approaching the human sequences such that all of the great apes have zero CCG tri-nucleotides. Interestingly, there appears to be an opposite trend in the codons of the amino acid Leucine, where one codon alone, CTG, has increased in frequency to increase the numbers of this residue. The mutation of CCG to CTG, immediately suggests the action of CpG methylation as an explanation. However, there is only a moderate, if any, increase in the frequency of CCA, which encodes Proline and may also result from mutation of a CpG site in a CCG codon.

										X lineage										Y lineage								
				EquusY	EquusX	MusX	OtolemurX	LemurX	Saimiri	Leontopithe	Ateles	Macaca	Colobus	Presbytis	Gorilla	Pan	Homo	Saimiri	Ateles	Macaca	Colobus	Hylobates	Pongo	Gorilla	Pan	Homo
	% of all codons		Total																							
	Genome*	AMEL**	codons	133	151	161	161	167	149	157	157	156	153	156	156	155	156	147	157	157	153	156	157	157	155	156
Pro (P)	**6.14**	**26.89**		**35**	**41**	**39**	**49**	**48**	**43**	**43**	**42**	**44**	**44**	**44**	**44**	**44**	**44**	**40**	**43**	**43**	**38**	**39**	**38**	**38**	**38**	**37**
ccc	2.01	9.88		9	13	15	22	23	15	15	16	16	18	16	16	16	16	11	15	17	12	14	13	15	15	14
cct	1.74	7.75		14	15	11	10	10	13	14	12	12	11	12	12	12	12	14	11	11	13	12	12	11	11	11
cca	1.68	7.27		11	8	11	10	9	11	10	11	11	11	11	11	11	11	12	14	12	13	12	13	12	12	12
ccg	0.71	1.99		1	5	2	7	6	4	4	3	5	4	5	5	5	5	3	3	3	0	1	0	0	0	0
Gln (Q)	**4.64**	**16.28**		**23**	**24**	**25**	**28**	**31**	**24**	**25**	**25**	**25**	**25**	**25**	**25**	**24**	**24**	**24**	**25**	**25**	**27**	**26**	**25**	**25**	**24**	**26**
cag	3.44	12.77		18	19	20	22	27	18	19	19	20	20	20	20	19	19	18	20	20	21	20	19	19	18	20
caa	1.2	3.51		5	5	5	6	4	6	6	6	5	5	5	5	5	5	6	5	5	6	6	6	6	6	6
Leu (L)	**10**	**8.67**		**12**	**16**	**14**	**13**	**14**	**12**	**11**	**13**	**13**	**12**	**12**	**12**	**12**	**12**	**11**	**13**	**14**	**16**	**16**	**15**	**15**	**15**	**16**
ctg	4.02	5.89		6	8	9	9	10	8	8	9	8	9	8	8	8	8	8	8	9	12	12	11	11	11	12
ctc	1.97	0.98		1	1	0	1	1	1	2	1	2	1	2	2	1	1	1	2	3	2	2	2	2	2	2
ctt	1.3	1.04		3	3	4	2	2	2	0	2	2	1	1	1	2	2	1	2	1	1	1	1	1	1	1
ttg	1.27	0.73		2	4	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
tta	0.74	0.03		0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cta	0.7	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
His (H)	**2.57**	**7.6**		**11**	**13**	**13**	**12**	**14**	**12**	**12**	**12**	**12**	**12**	**12**	**12**	**12**	**12**	**12**	**12**	**12**	**9**	**10**	**12**	**11**	**11**	**11**
cac	1.51	6.34		10	11	8	11	11	10	11	11	11	11	11	11	11	11	10	11	11	7	7	9	7	7	8
cat	1.06	1.26		1	2	5	1	3	2	1	1	1	1	1	1	1	1	2	1	1	2	3	3	4	4	3
Val (V)	**6.12**	**5.72**		**6**	**8**	**7**	**10**	**8**	**9**	**10**	**10**	**9**	**8**	**10**	**9**	**8**	**9**	**11**	**11**	**10**	**7**	**7**	**9**	**10**	**9**	**9**
gtg	2.86	3.87		4	4	6	7	4	6	7	6	7	5	7	7	6	7	7	7	7	4	5	6	7	6	6
gtc	1.47	0.56		2	3	0	1	1	1	1	2	0	0	1	0	0	0	2	1	0	1	0	1	1	1	1
gtt	1.09	1.21		0	1	1	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2
gta	0.7	0.08		0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Ile (I)	**4.41**	**3.45**		**9**	**7**	**6**	**4**	**7**	**5**	**5**	**5**	**4**	**5**	**4**	**6**	**6**	**6**	**5**	**5**	**4**	**4**	**5**	**5**	**6**	**5**	**5**
atc	2.11	2.33		5	6	3	4	6	3	3	3	3	3	3	5	5	5	3	3	3	3	3	3	2	3	3
att	1.58	0.28		1	0	2	0	0	1	1	1	0	0	0	0	0	0	1	2	0	0	0	0	1	0	0
ata	0.72	0.84		3	1	1	0	1	1	1	1	1	2	1	1	1	1	1	0	1	1	2	2	3	2	2

Table IV.5 - Frequencies of common amino acids and their codons in each AMEL sequence.

The fifth most common amino acid in AMEL is Methionine, but, with just one codon (ATG) and no distinguishable difference in X or Y lineages, this was excluded from the table. Codons are sorted according to their frequency in the human genome. * The genomic codon %s are for human and come from http://www.kazusa.or.jp/codon/ . **AMEL codon frequencies are an average across all codons and all sequences in the alignment.

IV.3.6 Mutation analysis

Because of the bias in GC->AT mutations over AT->GC mutations found previously in several Y linked genes (Eyre-Walker 1993), such mutations were assigned to specific branches of the phylogeny relating these sequences. Using alignment A1, and the tree shown in Figure IV.4, the programme *baseml* (Yang 1997) assigned mutations to specific branches on the tree. The output of this programme included ancestral sequences for each node of the tree and the directions of the changes along each branch. For the anthropoid primates (not lemurs), differences were then scored as being AT->GC or GC->AT for both the X and Y lineages. For the X lineage, the numbers of AT->GC and GC->AT changes along the entire alignment were 55 and 60 respectively (5 and 6 in the coding region). For the Y lineage, the numbers of changes were 95 and 121, respectively (10 and 22 in the coding region). Whilst the total number of mutations in the Y lineage was greater (see Chapter I), the proportions of AT->GC and GC->AT mutations were not significantly different between X and Y either for all sites (Fisher's exact test (1-tail), P=0.290), or for coding sites only (P=0.394).

IV.3.7 Characterisation of parsimoniously informative sites

Parsimoniously informative sites along alignment A2 (both coding and non-coding) were scored for shared-derived characters (changed since divergence from ancestors) inclusive of the putative Y sequences from

Macaca and Ateles (M10 APb and A09 AMc, respectively). If there had been recombination between X and Y sequences around exon 6 in either of these species, we might expect mutations derived in the X lineage to be present on the gametologous Y sequence more often than mutations that were shared derived with the Y sequences from other primates.

Table IV.6 lists the sites that were scored for the inclusion of sequence A09 AMc into X only, Y only or New World only lineages. Some sites, such as #60 (see appendix), did not easily fit into any sensible phylogenetic categorisation and were labelled as 'mixed'. Incidentally, sites 60 and 61 formed what may have been at one time a CpG site, though this is far from certain; if it was, it was the only one detectable in intron 5 (compared to 12 in the exons).

Table IV.6 – shared derived substitutions of sequence A09_AMc Site numbers correspond to position along alignment A2 in the Appendix *denotes a CpG site

Region	All New World	X only	Y only	Mixed
Non-coding	-	66	117, 123, 124, 136	60*, 106, 184, 200, 272
Coding	532, 606	-	45*	6, 34, 425, 509*, 773, 829

Table IV.7 shows the pattern for the inclusion of M10 APb from Macaca (as this is an Old World sequence it was not scored for inclusion in a New World lineage).

Table IV.7 – shared derived substitutions of sequence M10_APb Site numbers correspond to position along alignment A2 in Appendix 2 *denotes a CpG site

Region	X only	Y only	Mixed
Non-coding	0	117, 123, 124, 130, 136	60, 106, 184
Coding	792, 810	39, 45*, 727*	6, 34, 386, 779, 829

Of the twelve pairs of identifiable CG di-nucleotide sites along the coding alignment A3, all twelve have mutated in at least one lineage (including the ancestors). Where mutations occurred after the divergence of lemurs from the anthropoid apes (Apes and NW and OW monkeys), there were 12 mutations (at individual sites) that could be ascribed to an X sequence, a Y sequence or both independently. The respective numbers of mutations in these lineages were 0, 10 and 2. Of the ten Y linked mutations, six were shared by ape Y sequences alone or with C71 Y to the exclusion of both M10 APb and A09 AMc. Five of the 12 changes caused non-synonymous amino acid replacements, two from GTG (V) -> ATG (M) and three from CCG (P) to CTG (L). The pattern of replacements, both in the sequences involved and the amino acid changes, matches the pattern seen for Proline and Leucine in Table IV.5.

It was also observed that all along alignment A2, the two New World Y sequences from Saimiri and Ateles, shared many derived substitutions specific to these two sequences. Interestingly, the same can be said for the Y sequences from the OW monkey, Colobus (C71) and the ape, Hylobates (G72) which shared 9 specific derived substitutions as well as several substitutions which were shared with the other ape Y sequences, though 5 of these were at the CpG sites mentioned above. The numbers of shared derived sites that grouped each of these sequences with the higher apes but excluded the other were 2 and 3 for the inclusion of Colobus or Hylobates, respectively. Contrary to accepted relationships of Old World monkeys and apes, this data better supported Colobus as an ape, in a clade with Hylobates as per Figure IV.5. Unfortunately, no PCR product was obtained from Hylobates for AMELX with which to check the phylogenetic relationship of the X chromosomes of these two species. Equally, no Y sequence was obtained from the Colobus sample in the earlier study of SMCY, UTY and USP9Y (see Chapter III). The only other gene for which both samples gave sequence was UTX, which, amongst the apes and Old World monkeys is so conserved that no nodes in the tree can be resolved with any confidence (data not shown).

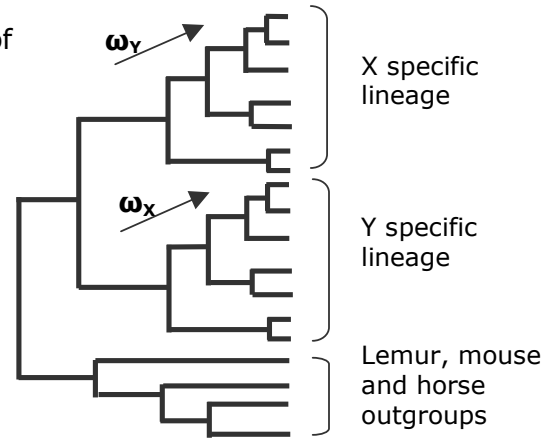
IV.3.8 Maximum likelihood tests

The models of evolution implemented to test the heterogeneity in ω, use a single phylogenetic tree and assume zero recombination. Their performance and reliability with a recombinant dataset are likely to give a positive result erroneously as the substitutions shared by misplaced taxa

resemble multiple mutations at one site (Anisimova, Nielsen, and Yang 2003).

Given the incongruencies between the trees based on different alignment methods and the apparent difference in rate of evolution between introns and exons, the maximum likelihood estimations were run initially using the tree in Figure IV.4 (based on alignment A1). Species not already present in this tree but present in alignment A2 were added to the tree according to the phylogeny given in Purvis (1995), which contradicted some of the trees generated from the amelogenin sequence. Sequences A09 AMc and M10 APb (from Ateles and Macaca), were assumed to be Y linked (as per Figure IV.4). The results of the maximum likelihood estimates and ratio tests are given in Table IV.8.

Figure IV.12 – Schematic of tree used for the likelihood ratio tests and different lineages that are tested for heterogeneity in ω.



Firstly, a branch model, allowing a different ω ratio in the Y lineage (see Figure IV.12) was tested against the base model of one ω for the entire tree (Table IV.8, Run 1). The LRT test of these models showed that the more parameter rich, two- ω model fit the data significantly better than the one- ω model. The ω ratios estimated for the Y lineage and the rest of

the sequences were 1.33 and 0.73, respectively. Because of the ω_Y >1.0, the next model (Run 2) tested for heterogeneity in ω among the sites in the sequence. However, as models that simultaneously estimate differences between branches and between sites (branch-site models) have been shown to be unreliable (Zhang 2004), the Y lineage (Figure IV.12) was removed from the overall tree and tested alone. The drawback of this method is that information would be lost from substitutions that occurred between the X and Y lineages but were shared by all Y sequences. The test between model M0 and model M3 is a test of whether the codon sites fit better under one ω ratio or under more (in this case just two), respectively. Again, the result was a significantly better fit for the less general model of heterogeneity in ω. The parameter estimates of 0.22 and 5.16 (30% of sites) are suggestive of a combination of sites under both negative and positive selection within the same gene. To better test for the presence of positive selection, models M14 and M15 were used (Run 3) to ascertain if, in addition to the constrained sites, a separate class of ω greater than one (M15) was a better fit to the data than an additional class of sites with ω equal to one (M14). Whilst model M14 estimated that 71% of sites in the Y lineage were undergoing neutral evolution (ω=1), M15 gave a better fit to the data and estimated a significant proportion of sites with a greater rate of non-synonymous than synonymous substitution (ω=5.50, 26% of sites).

In the initial run of models 14 and 15 on the Y only tree, 37 sites (26%) fell into the positively selected class of sites. Of these, only 5 had a posterior probability greater than 0.95 of being positively selected. When

the analysis was re-run with the Colobus and Hylobates grouped according to the sequences (i.e. forming a sister clade to the apes, with Macaca outside), the results of significantly different branch rates (Run 4), and of significant positive selection within the Y tree (Run 5), remained. However, the proportion of sites under positive selection dropped to 1.5%, all in the list from the previous analysis. Interestingly, the only two codon sites to retain greater than 0.95 posterior probability, were consecutive. They correspond to the six nucleotide sites 483-488 in alignment A2 (see Appendix). Even allowing for the clustering of Colobus with Hylobates, these six sites have received 7 base changes in the Y lineage alone, and all at first and second codon positions. However, site 483, which received 3 of the substitutions, forms a CpG site with the preceding nucleotide.

Following the removal of the Y lineage from the tree of amelogenin sequences, the anthropoid X lineage (Figure IV.12) was then tested for a different ω ratio relative to the rest of the remaining sequences (including horse, mouse and lemurs). The parameters estimated were similar to those from the test of the Y lineage in that the X sequences had an ω ratio greater than one and ω for the remainder of the tree was below one (Run 6). However, the difference in likelihoods in this case was not significant (P=0.06) and the null hypothesis of one ω along all branches in the tree cannot be rejected. Due to the marginal value of this test, the M0/M3 and M14/M15 tests were also made on the X specific lineage. Surprisingly, positive selection was also indicated to have driven changes in the X lineage. Both the M0/M3 (Run 7) and the M14/M15 (Run 8) tests

predicted 2% of sites with an ω ratio much larger than unity. Model M15 gave a posterior probability of selection for just two codon sites in the X lineage, both with probability greater than 0.95. Interestingly, one of these sites is the same as the second site predicted for the Y lineage and corresponds to nucleotides 486-488 in alignment A2 (see appendix). There are 2 or 3 changes (depending on the phylogeny within the Old World monkeys), all at the first site. The second positively selected site covers nucleotides 811-813 in alignment A2 and is either a recurrent mutation of G to T at site 811 or a problem in the phylogeny. The Colobus X sequence shares the ancestral G state with the New World monkeys, but the other old world monkeys and the apes have the derived T state. As Colobus and Presbytis (Hanuman langur, P70) should form a clade within the Old World monkeys (Purvis 1995), the relationships at this site are again incongruent with the phylogeny. However, no phylogeny would align the sequences such that there was no recurrent mutation between the two positively selected sites predicted in the X lineage. Either there is recurrent mutation or recombination. With the latter there must also be independent sorting of ancestral polymorphism amongst the different lineages. In addition, underneath the signal of positive selection, all other sites in the X specific alignment have been estimated an ω ratio around one. This is a reflection of the fact that most sites in the X lineage are invariable for both non-synonymous and synonymous changes.

Table IV.8 – Parameter estimates and Likelihood Ratio Tests of models of evolution on coding sequence in alignment A3.

[¥] $2\Delta L$ is double the difference in log-likelihoods of the two models. [†] p_0 and p_1 are the proportions of sites with ω_0 and ω_1 respectively. [§] the p-value of the M14/M15 comparison is half that of the χ^2 test (Swanson, Nielsen, and Yang 2003)3). [θ] Sequences from Colobus & Hylobates brought together into a clade in the phylogenetic tree. [λ] dof = degrees of freedom

Type/seqs	Model A	Model B	Parameters	$2\Delta L$[¥]	p-value (dof[λ])
1. Branch/ All sequences	ω_0 for all branches	ω_Y for Y lineage. ω_B for all else	A: ω_0=0.941 B: ω_B= 0.726 ω_Y= 1.331	4.404 (1.d.f.)	0.036 (1)
2. Sites/ Y lineage only	M0 ω_0 estimated for all sites	M3 2 site classes: ω_0, ω_1 (p_0, p_1)[†]	A: ω_0 = 1.510 B: ω_0 = 0.224 ω_1 = 5.155 p_1 = 0.303	27.378 (2.d.f.)	<0.001 (2)
3. Sites/ Y lineage only	M14 $\omega_0 = \beta$ $0< \beta<1$ & ω_1=1 (p_1)	M15 $\omega_0 = \beta$ $0< \beta<1$ & ω_1 >1 (p_1)	A: p_1 = 0.710 B: ω_1 = 5.501 p_1 = 0.260	26.604 (2.d.f.)	<0.001 [§] (1)
4. Branch/ All sequences (C71 & G72 rearranged)[θ]	ω_0 for all branches	ω_Y for Y lineage. ω_B for all else	A: ω_0=0.916 B: ω_B= 0.725 ω_Y= 1.324	3.939 (2.d.f.)	0.047 (1)
5. Sites/ Y lineage only (C71 & G72 rearranged) [θ]	M14 $\omega_0 = \beta$ $0< \beta<1$ & ω_1=1 (p_1)	M15 $\omega_0 = \beta$ $0< \beta<1$ & ω_1 >1 (p_1)	A: p_1 = 0.880 B: ω_1 = 4.632 p_1 = 0.015	10.80 (2.d.f.)	<0.001 (1)
6. Branch/ X lineage after Y lineage removed	ω_0 for all branches	ω_X for X lineage. ω_B for all else	A: ω_0= 0.710 B: ω_B= 0.609 ω_X= 1.552	3.317 (1.d.f.)	0.069 (1)
7. Sites/ X lineage only	M0 ω_0 estimated for all sites	M3 2 site classes: ω_0, ω_1 (p_0, p_1)[†]	A: ω_0 = 1.650 B: ω_0 = 1.16 ω_1 = 62.80 p_1 = 0.020	9.208 (2.d.f.)	0.0012 (2)
8. Sites/ X lineage only	M14 $\omega_0 = \beta$ $0< \beta<1$ & ω_1=1 (p_1)	M15 $\omega_0 = \beta$ $0< \beta<1$ & ω_1 >1 (p_1)	A: p_1 = 1.00 B: ω_1 = 48.66 p_1 = 0.020	10.45 (2.d.f.)	<0.001 (1)

IV.4 Discussion

This study set out to describe the rate and mode of molecular evolution within a mammalian gene since it had become discretely X and Y linked. The linkage of individual sequences within the primates should have been easy to determine based on their homology to human X or Y sequences. The boundary of the PAR and the non-recombining region is well described (Iwase et al. 2003) and, as the old strata 3-4 PAR boundary falls in intron 2 of this gene, there should have been a single, best phylogenetic tree relating the sequences from exons 3 to 6. By filling out the amelogenin Y tree for a range of primates, this study should have given an insight into the rate of accumulation of non-synonymous substitutions in a gene, which finds itself confined to the male-sex, on a chromosome that does not recombine. What went wrong?

IV.4.1 Difficult family trees

The first indication of an unexpected pattern in the new amelogenin sequences came when assembling the 840bp alignment (A2) of exons 5 and 6 and intron 5 following the first round of sequencing. More than a few sites in the relatively conserved coding region did not fit the accepted relationship between the species and chromosomes sampled. This was borne out in the phylogenetic trees of the region which featured the New World monkey sequences falling outside of the other X and Y of the Old World monkeys and apes (Figure 5). According to Iwase et al. (2003),

amelogenin split into distinct X and Y forms prior to the divergence of Old and New World monkeys (30-55 MYA).

Specifically, sequences A09 AMc (Ateles) and M10 APb (Macaca) were placed ambiguously. As all the Macaca and Ateles sequences had come from longer PCR products than other sequences, it was possible to carry out additional sequencing (from exon 3 back towards exon 6) to add information and phylogenetic resolution to the alignment of sequences. The agreement of the tree from the longer alignment, A1, with the accepted phylogeny of these taxa, confirmed A09 AMc and M10 APb as homologues of the human Y.

The different trees generated by the sub-alignments had to be explained. Whilst the lower divergence of exons (and shorter sequence) could be reasonably expected to reduce phylogenetic signal, the consistent attraction of M10 APb toward the X sequences and more significantly of both A09 AMc and the published Saimiri Y sequence toward the New World X sequences, suggested a more systematic failure of the assumptions made in creating a phylogenetic tree. Both recombination (gene conversion) and mutational bias were investigated as additional factors complicating the phylogenies of these sequences. The presence of either would violate the assumptions of the maximum likelihood techniques.

IV.4.2 Recombination: getting back with the X?

One possible reason why some part of a Y linked sequence would look more like an X sequence, would be gene conversion between the two chromosomes within a species. Whilst Iwase et al. (2003) had shown that the Y sequences from Squirrel monkey, Human and Chimpanzee were more closely related to each other than to their respective X chromosomes in the 3' direction from the old PAR boundary, their sequence coverage of Squirrel monkey only went as far as exon 6 where X and Y divergence bottomed out. The independent movement of the PAR in other mammalian lineages and the apparent difference in range of the boundary, left open the possibility that, for a time, the boundary had lain a little further downstream, and that recombination, perhaps in some lineages only, continued. In addition to this, gene conversion of the genes ZFX and ZFY has been detected in a phylogeny of cats (Pecon Slattery, Sanner-Wachter, and O'Brien 2000), so it may not be an uncommon event.

The four-gamete test revealed a signal of recombination both across the entire data set and amongst the X and Y lineages separately. The result from the Y lineage especially, seemed to be at odds with our understanding of evolution around the old PAR boundary. Furthermore, using just NW monkey sequences, it appeared that recombination had occurred between X and Y.

The programme TOPALi was used to search for a signal of recombination using three different methods. All three detected recombination according to their respective statistical thresholds and in each case the point of recombination (or the strongest signal in the case of the Dss method), lay 100bp upstream of exon 6. Immediately, this seemed to explain the differences in the phylogenetic trees. However, a further analysis using four sequences that had been consistently placed between the trees also showed signs of recombination, and in the same location: just before exon 6 (Figure 10).

The Dss method used in TOPALi uses trees generated from a distance matrix. Whilst the authors claim that there is some account taken of it, differences in rates of evolution between sequences is a common cause of failure of distance measures to find the correct topology (Milne et al. 2004). The other two methods though, HMM and PDM, rely only on the support for topology of the sequences and should cope better with rate variation. However, Posada and Crandall (2001) showed by simulation that all types of recombination detection methods might fail when there is extreme rate variation in different regions of a sequence. Additionally, even parsimony may fail to find the 'true' topology when the same character evolves multiple times in different lineages. Accordingly, both substitution rate variation and the possibility of homoplasy were investigated at the fine scale around exon 6.

IV.4.3 The paucity of frameshift indels and stop codons

The simplest analysis into the substitutions occurring within the mammalian amelogenins was the count of indels. As expected, indels were less common in exons than introns. The only primate frameshift indel occurred in the Y lineage of the NW monkeys in exon 4. As this exon is spliced from most human X linked transcripts and had also picked up frameshift indels in both horse X and Y sequences, it is perfectly possible that the exon is not true coding sequence in AMELY of the NW monkeys.

None of the new, or previously published amelogenin sequences contained stop codons (TGA, TAA or TAG); a fact, which may not be so revealing in light of the exceptionally biased %GC nucleotide composition. If there are any true pseudogenes in this data set, they may not have any stop codons yet because there are few codons that are a single mutation away from becoming a stop codon. The five to none ratio of in-frame to frameshift indels in exon 6 of the primates however, is suggestive that selective constraint is still strong enough here to weed out really destructive mutations.

IV.4.4 Mutation bias

The analysis of amino acid and base compositions revealed a potential source of mutation bias involving CpG sites. These are dinucleotides running 5’-C-G-3’ which, when the cytosine becomes methylated, stand

an increased chance of mutation by the hydrolytic deamination of 5-methylcytosine to thiamine. In the human genome, and in most mammals these sites mutate at an order of magnitude greater frequency than normal (Bird 1980). Table IV.5 suggested an increase in the mutation rate from Proline codon CCG to the Leucine codon CTG in the Y lineage. The study of bias in mutations assigned to X and Y lineages, revealed a non-significant excess of mutations from GC->AT along the Y relative to X, in the same direction as the significant trend found by Eyre-Walker (1993).

The study of derived characters shared by A09 AMc and M10 APb at the parsimoniously informative sites seemed to be little affected by hyper-mutating CpG sites. Two such sites lent support to the inclusion of these sequences in the Y clade and two did not support any particular clade unambiguously. Whilst intron 5 seemed to favour the Y clade for both, in exon 6, support for inclusion of these sequences in the X or Y clades was equivocal.

The analysis of the CpG sites clearly supports the exclusion of the Ateles and Macaca Y sequences within the coding region, but this is based on the fact that many of the CpG mutations seem to have happened in the apes and in the Colobus Y sequence. However, the multiple changes that seem to have occurred between just the ape sequences at some of the CpG sites are enough to support the hypothesis that there was a decrease in constraint acting upon them. Sites 366, 405, 483, 581, 749 & 829 (see appendix) are all CpG sites that look like homoplasy within the Y lineage.

IV.4.5 Models of evolution

The relatively large number of shared derived sites between the *Colobus spp.* (Colobus monkey) and *Hylobates lar* (Lar gibbon) sequences suggests that the accepted phylogeny of primates (Purvis 1995), does not best explain the relationship of these two sequences. A large portion of the initial signal of positive evolution was due to the separation of these two species. Few studies give divergence estimates for both lineages using one method but Yoder and Yang (2000) estimated 30-40 MYA and 19-22 MYA for the divergence from the human lineage of OW monkeys (Colobus) and Hylobates, respectively. It seems very unlikely that polymorphic Y chromosomes persisted between these divergences. Unfortunately, with the lack of corroboration from other genes from the Colobus sample, the most likely explanation is that either the sample (from the Institute of Zoology, London) was mis-labelled, or there was contamination of PCR products by species not otherwise amplified. This latter option is unlikely because the sequence clusters with the apes, and different sequence was obtained from all the ape samples.

Accounting for the grouping of the Colobus and Hylobates sequences, the heterogeneity in ω found between the Y lineage and the rest of the tree, and between the codons of the Y lineage, would normally be interpreted as evidence of positive selection acting to inflate the rate of non-synonymous amino acid change at some sites within the gene. However, in this case the signal may come from the relaxation of selection. If

hyper-mutating sites had been under strong constraint not to alter in the past, then once constraint is reduced, these sites would show a preponderance to change in each lineage. Over an entire phylogenetic tree, such sites would appear to undergo a greater rate of evolutionary change than other, neutrally evolving sites. This is supported by the involvement of the most mutable positively selected site being a CpG site.

IV.4.6 Further work

Additional sequencing of intron 6 in the primates used in this study may solve most of the problems and unresolved issues discussed above. If intron 6 was found to have a different phylogenetic topology to that in Figure IV.4, then this may well be good evidence for genuine gene conversion in amelogenin during the evolution of anthropoid primates. If the phylogeny was consistent with Figure IV.4, then this would not rule out gene conversion of exon 6 but would better support a mutational rather than a recombinational explanation for the incongruence.

A foreboding final word comes from a recent study of the accepted forensic test for genetic maleness (Chang, Burgoyne, and Both 2003). Whilst the failure rate in Malaysian and Chinese men was between 0-0.6%, over 3% of Indian men tested, failed the test. Other Y linked markers confirmed that the men did possess Y chromosomes and that a

deletion of the region containing the test marker was segregating in the population. The marker used to test for maleness is AMELY[2].

[2] No mention was made of the state of the men's teeth.

V SEQUENCING OF SLX1 IN A SAMPLE OF *SILENE LATIFOLIA* AND *SILENE DIOICA*

Dave T. Gerrard

School of Biosciences,

The University of Birmingham,

Edgbaston, Birmingham, B15 2TT

V.1 Introduction

V.1.1 The Silene sex chromosomes

Silene latifolia (White campion)[†] and *Silene dioica* (Red campion) are two of six dioecious species in the large genus of around 700 hermaphroditic and gynodioecious Silene (Desfeux et al. 1996). Their recently evolved sex chromosomes are considered to offer one of the best opportunities to study the early evolution of sex chromosomes. As recombination of the Y chromosome ceased less than 20 million years ago (MYA) (Desfeux et al. 1996; Atanassov et al. 2001), it bears only some of the degenerative features characterising the more ancient sex chromosomes of mammals and Drosophila (see Chapter I).

S. latifolia has a pseudoautosomal region (PAR) where the short (p) arm of the X chromosome pairs with the long (q) arm of the Y chromosome (Lengerova et al. 2003). Moore et al. (2003) mapped the gene DD44X to the distal end of the Xq arm opposite to the PAR. The four genes detected so far that are shared by X and Y are, in order from the PAR: SlX1/SlY1, DD44X/ DD44Y, SlssX/ SlssY and SlX4/ SlY4. The X-Y divergences for the four genes are 1.7%, 7%, 8% & 16%, respectively. Filatov (2004, submitted) found that the homologues of the four X linked genes are also linked in the hermaphroditic *S. vulgaris,* though with SlssX and SlX4 in switched relative positions. These results suggest that the genetic content

[†] Previously *Melandrium album* & *Silene alba*

of the X chromosome has remained unchanged since the evolution of separate sexes and therefore since the evolution of the sex chromosomes from a pair of homologous autosomes. The conservation of the genetic content of the X is similar to the situation in mammals in which the X chromosomes of distantly related species (e.g. man vs. mouse) have a very similar set of genes and have not experienced the level of large-scale chromosomal rearrangements of other chromosomes. In addition, though there are effectively only three data points, the sequential increase in X-Y divergence moving along the chromosome (25.5cM separate SlX1 and SlX4) is reminiscent of the strata of the human X chromosome (Lahn and Page 1999) for which recombination ceased at different times.

Similar to the Y chromosomes of animals, the *S. latifolia* Y chromosome contains both a male determining factor (which suppresses carpel formation), and male fertility factors (Lebel-Hardenack et al. 2002). Matsunaga et al. (2003) found a Y linked MADS box gene in *S. latifolia* that shares closest homology with an autosomal gene, has no detectable X chromosomal homologue and is expressed more strongly in developing stamens than its autosomal paralogues. Though the Y chromosome is still largely euchromatic (Siroky, Castiglione, and Vyskot 1998), there are already signs that it is also starting to degenerate. The first sex-linked gene described in Silene, MROS3, had a degenerating Y linked homologue (Guttman and Charlesworth 1998). Obara et al. (2002) found that a gene which is divergent between X and Y in *S. latifolia*, was missing from the Y chromosome of *S.dioica*.

V.1.2 Introgression

In hybrids of *S. dioica* and *S. latifolia*, which are generally fertile (Taylor 1994), X and Y chromosomal sequence from both species has been found in single individuals (e.g. Filatov et al. 2001). In these cases, the molecular species assignment of the X linked sequence does not match the assignment based on phenotypic characters (D. Filatov, pers. com.). In other individuals, both sex chromosomes may be from one species whilst an autosomal linked gene is from the other, indicating that genes which introgress between species can be passed into the offspring of hybrids (e.g. Filatov et al. 2001). However, a conflict between the phenotypic classification of a plant and its Y linked sequence has yet to be observed; suggesting that the Y chromosomes are unable to introgress between species (D.Filatov, pers. com.). Introgression has been observed in the genes SlX1, CCLS37.1 (e.g. Filatov et al. 2001), SlX4 (Laporte, Filatov and Charlesworth, in press) and in DD44X (Filatov, pers. com.) from different European population samples.

V.1.3 Ancestral Polymorphisms

The maintenance of ancestral polymorphisms may be an alternative explanation for the appearance of *S. latifolia* like X sequences within *S. dioica* plants, or vice versa. The two species are very closely related and their speciation began within the last 20 million years (Atanassov et al. 2001) and probably well after the sex chromosomes evolved. Published

divergence between the X chromosomes of the two species is 1.4% and 1.3% for SlX1 and SlX4, respectively (Atanassov et al. 2001).

Though 20 million years is a long time for separate species to be sharing polymorphisms that were also polymorphic in their common ancestor, this figure is an upper estimate. Clark (1997) gives the mean time to loss of a polymorphism from one of a pair of sibling species as $1.7N_e$ generations, where N_e is the effective population size. However, this theoretical distribution has a long right-hand tail and in 5% of cases, the same neutral polymorphism may hang around in both populations for 3.8 N_e generations. The diversity estimates of autosomal and X linked *S. latifolia* genes (1-2%) (Filatov et al. 2001) are comparable with Drosophila, for which the effective population size is estimated in the hundreds of thousands or millions. Whilst *S. latifolia* can reproduce from seed in two months, they are perennial plants and have an average generation time of 2-3 years (D. Filatov, pers. com.). Using coalescent simulations and assuming no gene flow, Filatov et al. (2001) estimated the divergence of *S. latifolia* and *S. dioica* to have occurred between $2N_e$ and $4N_e$ generations ago. The different population sizes of X and Y linked loci means that polymorphisms will be lost relative faster from Y than X. Filatov et al. (2001) also found that diversity in SlY1 was much reduced relative to SlX1.

V.1.4 Linkage disequilibrium

It may be possible to distinguish between introgression and the maintenance of ancestral polymorphisms by analysing linkage disequilibrium along the X chromosome. If ancestral polymorphisms are still segregating at two loci in the populations of both species, then the alleles at each locus should be in linkage equilibrium with one another. However, if the entire X chromosome of one species is introgressing into the other species, then the alleles originating from one species will be found together more often than by chance.

Machado et al. (2002) developed this idea for several closely related species of Drosophila. When a chromosome passes from one population into another, then, depending on the degree of isolation between the populations, it may introduce polymorphisms that were exclusive to the donor population. Such polymorphisms now become shared between the two populations. However, the newly shared polymorphisms will be in negative linkage disequilibrium with polymorphisms that were already segregating exclusively in the recipient population. Conversely, the disequilibrium between different shared polymorphisms will tend to be positive because polymorphisms imported on the same chromosome will be linked. Machado et al. (2002) proposed subtracting a measure of the first kind of disequilibrium (negative under introgression) from the second (positive under introgression) to test for introgression. The strength of disequilibrium will depend mostly on the recombination rate between the two loci: the higher the recombination rate, the weaker the signal, as

associations between loci are broken down. Under the alternative model of ancestral polymorphisms, the newly arisen, exclusive polymorphisms within a population should appear dispersed along the chromosome. In this case, there should be little or no disequilibrium between shared ancestral and exclusively segregating polymorphisms.

The main objective of this study was to sequence two regions of a single X chromosome in a sample of *S. dioica* and *S. latifolia* plants and this will enable us to make the test described above.

V.1.5 The cessation of recombination between X and Y

Molecular phylogenies of X and Y linked genes in Silene show the sex chromosomes to have been present in the common ancestor of both these species, but appear to differ in estimating when the cessation of recombination occurred. For example, sequences from the SlY4 of *S. latifolia* are more closely related to the SlY4 of *S. dioica* than to homologous sequences on the X of the same species. In the case of SlXY4 the split of X and Y (now 15.5% diverged (Filatov and Charlesworth 2002)) appears to have happened well before the speciation (~2% between species). However, for SlXY1 the cessation of recombination was much closer to the time of speciation as shown by Figure V.1, which is based on a figure in Filatov and Charlesworth (2002). That diagram was produced without statistical support for the reliability of the tree's topology and X-Y divergence for both species is in the same range as that between species. Whether the *S. dioica* and *S. latifolia* SlY1 sequences are

truly homologous, or whether each is descended from a con-specific X sequence after speciation, is an open question. However, as long as there is no recombination between X and Y in the living populations, the exact history of X-Y divergence at this locus should not affect the planned analysis. Filatov et al. (2000) indeed found no shared polymorphisms between X and Y loci in *S. latifolia* and no signal of recombination amongst their sample of 14 Y chromosomes when there was a clear signal of genetic exchange amongst the X chromosomes. In this study, a larger sample of SlX1 and SlY1 sequences generated by D. Filatov and myself from both species will be analysed with this in mind.

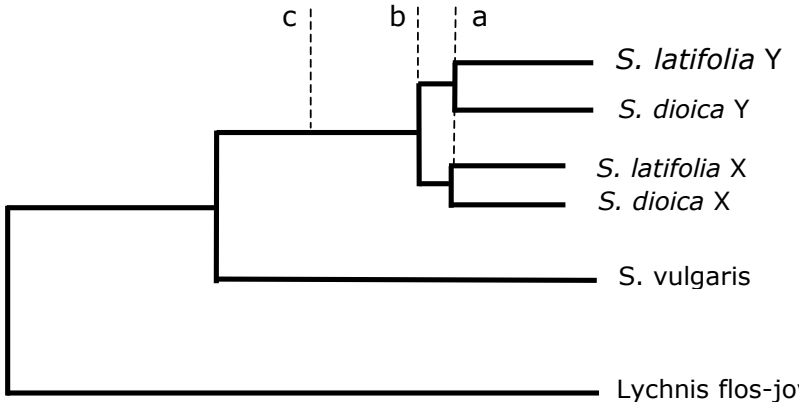


Figure V.1 – schematic of the relationships between SlX1 and SlY1 sequences in *S. latifolia* and *S. dioica*. Based on a neighbour joining tree in Filatov & Charlesworth, 2002. (a) marks the speciation of *S. dioica* and *S. latifolia*. (b) marks the cessation of recombination between SlX1 and SlY1. (c) marks the cessation of recombination between SlX4 and SlY4 which occurred at an earlier time.

V.2 Materials and Methods

The molecular study of Silene sex chromosomes is, like the sex chromosomes themselves, relatively new. Only a handful of sex-linked genes have been isolated. To measure linkage disequilibrium on a potentially introgressing chromosome, two *cis* sequences were required from an individual X chromosome in a range of individuals. The genes DD44X and SlX1 were chosen based on familiarity and their reasonably close linkage (~7.5cM, Filatov, submitted). As all individuals were males, the sequences from both genes would be from the same chromosome. Dr J. Ironside sequenced the gene DD44X from a sample of 17 *Silene latifolia* and 6 *Silene dioica* males. My part in the project was to sequence a 2kb portion of the gene SlX1 from those same individuals.

V.2.1 Collection of samples

The male plants were collected by Dr J Ironside and Dr D Filatov and were identified based on flower colour (White – *S. latifolia*; Pink – *S. dioica*).

Table V.1 lists the plant samples available for this study.

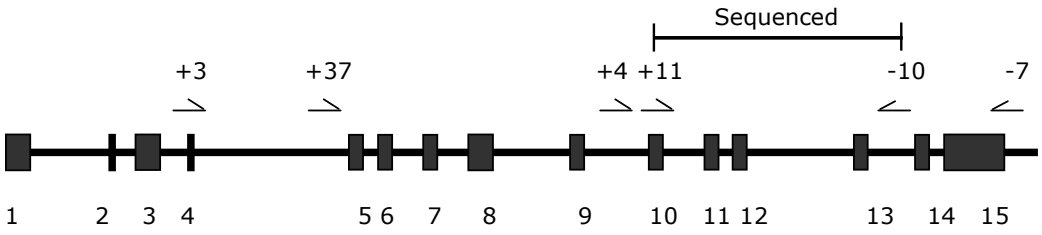


Figure V.2 – Exon structure of SlX1 spanning 8023bp.
Exons are numbered along the bottom. Main PCR primers are numbered along the top.

V.2.2 PCR & sequencing

A region spanning exon 10 to intron 13 of SlX1 (marked in Figure V.2) was sequenced from the same individuals that had been used for DD44X. The primers listed in Table V.2 were used, and the annealing sites of the PCR primers relative to the exons are described in Figure 1. PCR products were directly sequenced using the PCR primers and their homology to X or Y was established by a BLAST search (www.ncbi.nlm.nih.gov).

Table V.1 - List of samples and their place of origin

Lab code	Collector's code	Place of origin
S. latifolia		
Sa05	SaBC1	Alcoy, Valencia, Spain
Sa07	SaBR1	Alcoy, Valencia, Spain
Sa08	SaCB	Cluj botanic garden, Romania
Sa09	516	Sussex, England
Sa10	205	Dalkeith, Scotland
Sa12	524	Denmark
Sa13	SaMB3	Muntele Baisorii, Romania
Sa14	SaVL1	Valea Lerii, Romania
Sa15	SaMB2	Muntele Baisorii, Romania
Sa16	SaLo	Hampstead Heath, London, England
Sa18	SaWF	Wyre Forest, Worcestershire, England
Sa20	SaVM1	Dupa Plese, Romania
Sa29	SaSM1	Alcoy, Valencia, Spain
Sa33	SaBM	Baile, Romania
Sa34	SaMa	Malvern, Worcestershire, England
Sa35	SaGh1	Ghetari, Romania
Sa36	SaVI	Vidra, Romania
S. dioica		
Sd01	253	Blackford glen, Edinburgh, Scotland
Sd11	251	Blackford glen, Edinburgh, Scotland
Sd21	SdLo	Hampstead Heath, London
Sd23	278	North Berwick, Scotland
Sd26	SdWo	Worcester, Worcestershire, England
Sd28	SdMa	Malvern, Worcestershire, England

At this point, it was noted that Genbank contained two distinctly different *Silene latifolia* sequences for both SlX1 and SlY1 (four sequences in total), each pair of X and Y from a different study (Delichere et al. 1999; Filatov and Charlesworth 2002). Both the X and the Y sequence submitted by Filatov and Charlesworth feature a 250bp deletion in intron 12 relative to those submitted by Delichère et al.. This was described as a 150bp region by Filatov and Charlesworth and left out of their alignment (and apparently the Genbank submission) because its high AT% and multiple indels proved difficult to align. Whilst, the two SlY1 sequences (AY084037[†] (Filatov and Charlesworth 2002) and AJ310655 (Delichere et al. 1999)) share 99% identity outside of the large indel, the two SlX1 sequences (AY084036 (Filatov and Charlesworth 2002) and AJ310656 (Delichere et al. 1999)) share just 97% identity outside of the indel (Based on a two sequence BLAST alignment). On pure identity, these two X sequences are as similar to the Y sequences as they are to each other. Whilst, assignment of the Delichère et al. (1999) sequences was on the basis of a segregation analysis, the Filatov and Charlesworth sequences were assigned by identity (Filatov, pers com.). All four sequences, plus those from *S. dioica* X (AY084044) and Y (AY084045) were used to generate phylogenetic trees.

Due to repeated failures in PCR, not all products were generated using the same set of primers and this is recorded in the results. Sequencing primers are listed in order in Table V.2 and were each spaced around

[†] Genbank accession numbers

500bp apart along the locus. PCR was carried out as per Chapter III using the annealing temperatures listed in Table V.2. For the *S. latifolia* products, direct sequencing indicated that only one product was present (by a lack of heterozygous 'double-peak' sites in the chromatograph) and direct sequencing was used to complete these sequences. For *S. dioica*, initial sequence was weak or contained multiple products and these products were cloned. Sequencing and cloning were both carried out as in Chapter III.

Table V.2 – PCR & sequencing primers used, listed in order along the locus.
* (Filatov and Charlesworth, 2002) †Used in PCR

Primer	Sequence	Annealing temperature
SlXY1+3†	AGGCTCGTTCTCCCTTTGTG*	53.3ºC
SlXY1+37†	ATGTTAGCTGCTGTGTTGGAAAG	53.0ºC
SlXY1+4†	CTGGTTGCCACTTTCCACTTGC	58.3ºC
SlXY1+11†	AAGCTCACAATGCTGATCTTCACTG*	56.6ºC
SlXY1+42	TGCAGTGGTCACCGCATAATAG	55.2ºC
SlXY1-61	CATATGTTTAGCAAGCCATCTTCAG	54.2ºC
SlXY1+43	TGATCTAACTGCTCTGGGTGATC	52.6ºC
SlXY1-62	CCTCCGCTCTCTGTTTCCATC	55.2ºC
SlXY1+44	TAGGGACAAAGTGGTGGACTTCC	55.9ºC
SlXY1-10†	TCCAGCAGAGCTTGAACAGTC*	52.0ºC
SlX1-7†	ACTTGCAACGACTTCACTTTGAG*	53.0ºC

V.2.3 Additional sequences of SlX1 and SlY1

D. Filatov supplied an additional set of 74 sequences corresponding to the same region. These were X/Y pairs of sequences from 13 *S. dioica* and 26 *S. latifolia* individuals from around Europe. Each pair had been assigned to

X and Y by D. Filatov based on sequence identity to existing sequences. Four X sequences did not come with a partner from the Y.

V.2.4 Alignment and phylogenies

The new sequences were formed into contigs and aligned simultaneously using a Gap4 database (Staden, Beal, and Bonfield 2000) as in Chapter III. SlX1 and SlY1 sequences from *S. latifolia* and *S. dioica* and outgroup, autosomal-linked sequences from *S. vulgaris, S. conica* and *S. flos-jovis* (alt. *Lychnis flos-jovis*), which do not have sex chromosomes, were also retrieved from Genbank to help construct a phylogeny. The *S. dioica* sequences and all three of the outgroup sequences originated from the study by Filatov and Charlesworth (2002) and lacked the same 250bp region as detailed above. The range of pairwise divergences (Tamura and Nei 1993) between the *S.dioica* and *S. latifolia* sequences and the *S. conica*, *S.vulgaris* and *S.flos-jovis* out group sequences were 5.8-7.3%, 6.2-7.5% and 10.0-11.2%, respectively.

As the Genbank sequences were substantially longer than the region sequenced for the linkage disequilibrium study, they were formed into a long alignment by eye (4.8kb), which was used to check the topology of X and Y sequences in *S. dioica* and *S. latifolia*. This was performed in MEGA2 (Kumar et al. 2001) using both neighbour joining (NJ) and maximum parsimony methods, both with 1000 pseudo-replicates to provide bootstrap support. For the NJ tree, Tamura-Nei (1993) distances were used to account for multiple substitutions, transition/transversion bias and

differing base frequencies among the sequences. The NJ method was repeated with all gaps deleted from all sequences in the alignment to assess the impact of the missing 250bp of sequence from several Genbank sequences.

The new sequences produced in this study plus those donated by D. Filatov were then added to the long alignment. Due to the number of sequences, no consensus maximum parsimony trees were generated for the full alignment, which contained 98 unique sequences. The NJ method was applied as above to generate a phylogenetic tree with bootstrap support.

V.2.5 Tests for recombination

Filatov et al. (2000) found no shared polymorphisms between SlX1 and SlY1 in a population of *S. latifolia*. The pairwise divergence estimates between SlX1 and SlY1 suggest that recombination stopped quite recently (Atanassov et al. 2001). *S. dioica* and *S. latifolia* X and Y sequences from the 4.8kb alignment were tested for recombination using the four gamete test (Hudson and Kaplan 1985) implemented in DNAsp (Rozas and Rozas 1999). The same test was applied to each class of sequences: *S. latifolia X, S. latifolia Y, S. dioica X* and *S. dioica Y*.

The sequence from Sa 14 was difficult to classify as X or Y after BLAST searching or by placement in the phylogenetic tree of sequences. The Hidden Markov Method (HMM) implemented in TOPALi (Milne et al. 2004)

was used to measure the support for the different possible topologies between this sequence and the *S. latifolia Y, S. latifolia X* and *S. dioica Y* sequences. This method is described in Chapter IV and again, the default run parameters were used.

V.2.6 Measuring linkage disequilibrium

The combined data set of DD44X and SlX1 sequences will be analysed for linkage disequilibrium by Dr J. Ironside, now at Aberystwyth University.

V.3 Results

V.3.1 PCR & Sequencing

The success of PCR and sequencing efforts for each individual are outlined in Table V.3. The lengths of sequence in the table are estimated from bands seen on the agarose gel. 'Fail' means that no band was visible on a gel where at least one other individual gave a positive band in the same reaction. 'Weak' means that there was a band but that I was unable to obtain sequence direct from the band or clone it successfully. (s) denotes that a PCR product was used to generate sequence. (Y) denotes that the PCR product when partially sequenced, returned a top BLAST result as Y chromosomal. Initial sequence from individual Sa14, match both X and Y sequences in Genbank equally well, and was sequenced fully to determine its origin (see below).

V.3.2 Phylogenies

Figures V.3a, b & c illustrate the relationship of the X and Y Silene sequences in Genbank. The sequences are 4-4.5kb in length and the alignment (made by eye) spans 4.8kb due to the presence of many indels. Figure V.3a was created using the neighbour joining method with all sites included (gaps only excluded during each pairwise comparison). Whilst the two *S. latifolia* Y sequences cluster, the remaining *S. dioica* Y sequence and the X sequences from both species are less well resolved.

Table V.3 – Success and failure to isolate SlX1 from the samples
See text for explanation. *Probably a Y sequence – see text.

Sample	SlXY1+3 SlX-7 5.0kb	S.dio +4 SlX-7 2.6kb	SlXY1+11 SlX-7 2.2kb	SlXY1+42 SlX-7 1.6kb	S.dio +4 SlXY1-10 2.1kb	SlXY1+11 SlXY1-10 1.7kb	SlXY1+37 SlXY1-10 3.6kb
Silene latifolia							
Sa05	Fail	-	2.2kb (s)	1.5kb	2kb (Y)	-	-
Sa07	Fail	-	2.2kb (s)	1.5kb	-	-	-
Sa08	Fail	-	Fail	Weak	Fail	Fail	-
Sa09	5kb	-	2.2kb (s)	1.5kb	-	-	-
Sa10	Fail	-	2.2kb (s)	1.5kb	-	-	-
Sa12	Fail	-	Weak	Fail	Fail	Fail	-
Sa13	Fail	-	Fail	-	Fail	Fail	-
Sa14	5kb	-	2.2kb (s)*	-	-	-	-
Sa15	Fail	-	2.2kb (s)	-	-	-	-
Sa16	5kb	-	2.2kb (s)	-	-	-	-
Sa18	5kb	-	2.2kb (s)	-	-	-	-
Sa20	Fail	-	Fail	-	2kb (Y)	Weak	-
Sa29	Weak	-	Weak	-	2kb (Y)	-	-
Sa33	Fail	-	Fail	-	2kb (Y)	Weak	-
Sa34	5kb	-	2.2kb (s)	-	-	-	-
Sa35	Fail	-	Fail	-	Fail	Weak	-
Sa36	-	-	2.2kb (s)	-	-	-	-
Silene dioica							
Sd01	Fail	-	Fail	-	2kb (Y)	Fail	Fail
Sd11	-	-	Fail	-	Fail	1.6kb (s)	2.7kb (Y)
Sd21	-	-	2.2kb (s)	-	Fail	Fail	2.7kb (Y)
Sd23	-	2.7kb (s)	Fail	-	Weak	-	2.7kb (Y)
Sd26	-	-	2.2kb (s)	-	1.5kb	-	2.7kb (Y)
Sd28	-	-	2.2kb (s)	-	Weak	-	2.7kb (Y)

Interestingly, when gaps are removed from all sequences, the topology of the latter sequences changes and the bootstrap support increases (Figure V.3b). Under this tree, the *S. dioica* Y sequence joins the *S. latifolia* Y sequences, though there has clearly been a greater rate of evolution in the latter. The X sequences from both species do not form a separate clade, as would be expected if recombination had ceased between X and Y before speciation and there had been no gene flow between the species since. The increase in support for the nodes of this tree, even though it

uses less information, may be due to the removal of the indel mentioned above, which seems to confound the sequence relationships. Figure V.3c, which gives the topology under the parsimony method, lends less support to the Y clade, and fails to resolve the X sequences at all with bootstrap support less than 50%.

Figure V.3 – Phylogenetic trees relating SlXY1 sequences from Genbank. *S. dioica* sequences prefixed with 'd'. (a) Neighbour joining tree using all sites. (b) Neighbour joining tree with gaps deleted. (c) Parsimony tree. Numbers at the nodes are bootstrap support. Scale is substitutions per site in NJ trees a & b.

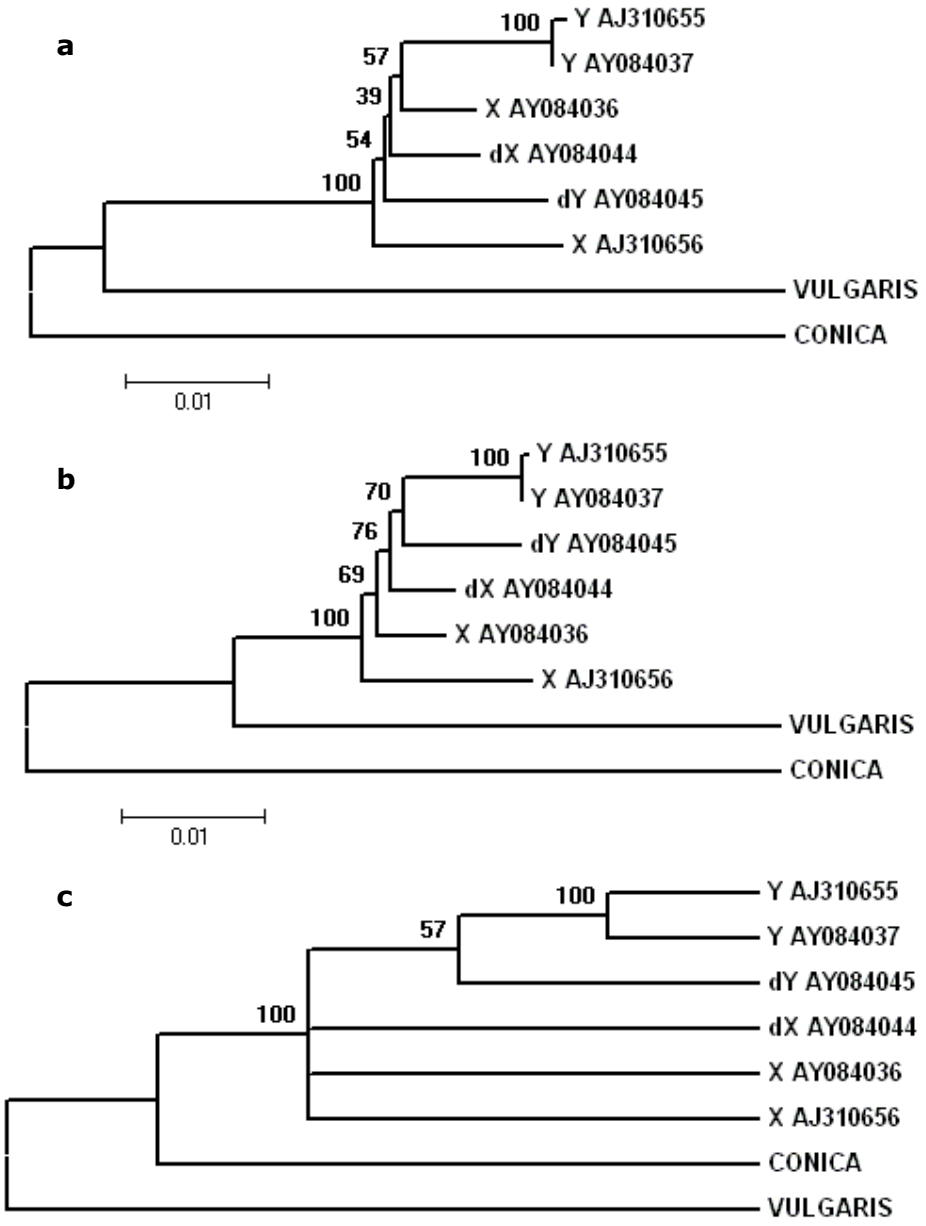


Figure V.4 represents the bootstrap consensus neighbour joining tree relating the sequences generated for this study. The new *S. latifolia* sequences obtained in this study are prefixed by SlX1. The *S. dioica* sequences begin with the individual's code followed by Y1 or X1 and then a clone number. The main aspect of the tree is that, within the *S. latifolia* and *S. dioica* clade, there is no resolution at the base of five strong sub-clades. These comprise two groups containing SlX1-like sequences, *S. dioica* X and Y clades and a mixed group containing the two Genbank *S. latifolia* Y sequences, one of the Genbank X sequences and the sequence from individual Sa14, which seems to fall between the two types. Two other inferences about the new sequences are that the sequence from Sd11 looks very like a Y sequence and that the sequence from Sd23 looks like a *S. latifolia* sequence. Checking back, the collector of this individual wrote that this individual was suspected to be a hybrid.

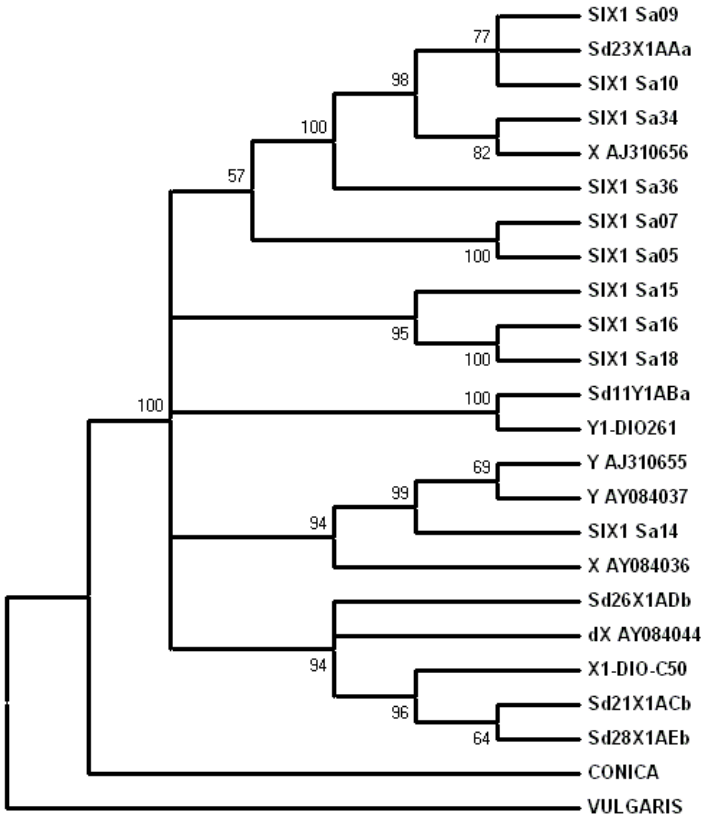


Figure V.4 – Bootstrap consensus neighbour joining tree of new sequences with other sequences from genbank. *S. latifolia* sequences begin SlX1 and end with an individual code. *S. dioica* sequences, which were cloned, begin with an individual code followed by X1 or Y1 and then the clone identifier.

Figure V.5 shows the NJ tree generated using sites shared between the newly generated sequences and those from Genbank and D. Filatov. The tree generally separates sequences according to whether they are X or Y and whether they are from *S. dioica* or *S. latifolia*. The part of the tree showing *S. latifolia* X and Y sequences is reproduced in Figure V.6. Likewise, the *S. dioica* region of the tree is reproduced in Figure V.7. In all three trees, nodes with bootstrap supports of less than 50% have not been labelled.

Figure V.6, comprising mostly the *S. latifolia* sequences from the large phylogeny, also fails to resolve the origination of the *S. latifolia* Y sequences. The nodes around the base of this sub-tree all have less than 50% bootstrap support. However, each pair of sequences donated by D. Filatov has split, with the X and Y sequences clustering separately. Interestingly, the X clade has deeper branches than the Y,

Figure V.5
NJ tree of all Silene sequences from this study and from D. Filatov

representing the difference in diversity between the two sex chromosomal gametologues. By comparison, all of the Y sequences, which are taken from the same sample of individuals as the X sequences, are very similar to one another. This pattern was described by Filatov et al. (2000), when they noted a 20-fold difference in diversity between X and Y. The newly sequenced individuals (beginning SlX1) are evenly spread throughout the X portion of the tree. Again, the sequence from the hybrid Sd23 falls deep inside the *S. latifolia* X clade. The sequence from Sa14 now clusters strongly with the Y-clade but still on the outside. A second sequence from this individual may help to determine which chromosome the sequence came from but at present, it more closely resembles Y than X and cannot be included in the linkage disequilibrium study. It is also interesting that several X sequences, including one of those from Genbank (X AY084036), cluster nearby Sa14 and the Y clade. Though bootstrap support in this region is low, these sequences may be closer to the ancestor of *S. latifolia* Y sequences than other X chromosomes in the sample.

The sub-tree representing the *S. dioica* X and Y sequences (Figure V.7), also shows a perfect split between X and Y sequences and again the branches in the Y clade are shorter than in the X clade. Though the latter feature is less pronounced than for *S. latifolia* (Figure V.6), in line with the finding of Filatov et al. (2001). In addition, the grouping of all the *S. dioica* X and Y sequences away from all *S. latifolia* sequences is well supported. This is surprising for two reasons. Firstly, this would suggest that the *S. dioica* Y linked sequences are descended from a con-specific X sequence and that the *S. latifolia* Y sequences are not their closest

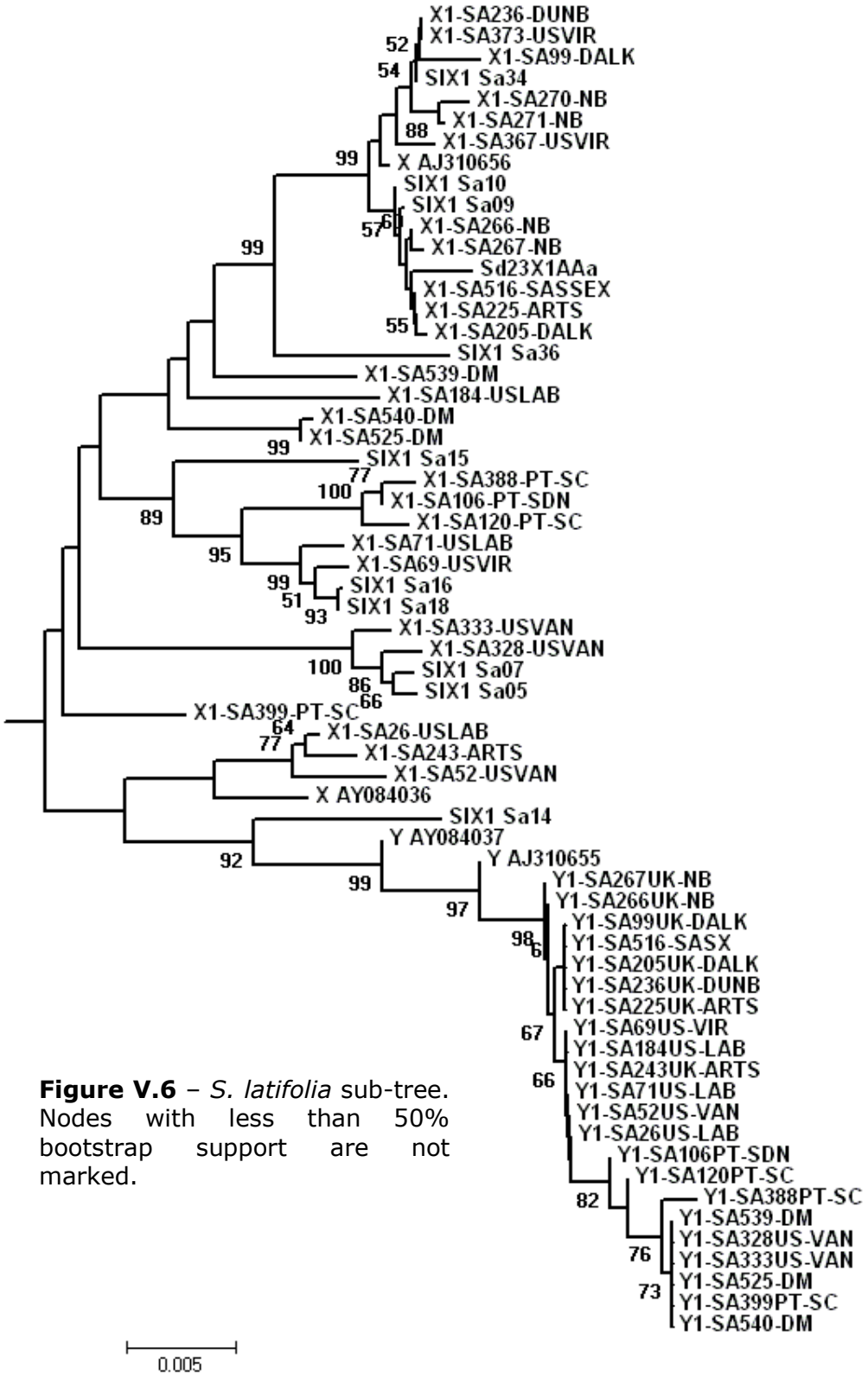


Figure V.6 – *S. latifolia* sub-tree. Nodes with less than 50% bootstrap support are not marked.

relative. Recombination between X and Y ceased independently at this locus. Secondly, this topology forces the clade of *S. latifolia* X sequences

outside of the *S. dioica* X and Y clade. The level or perhaps the direction of recent introgression between the X chromosomes of the two species must have been such that the signal of *S. dioica* X and Y homology in this region has not been over-written.

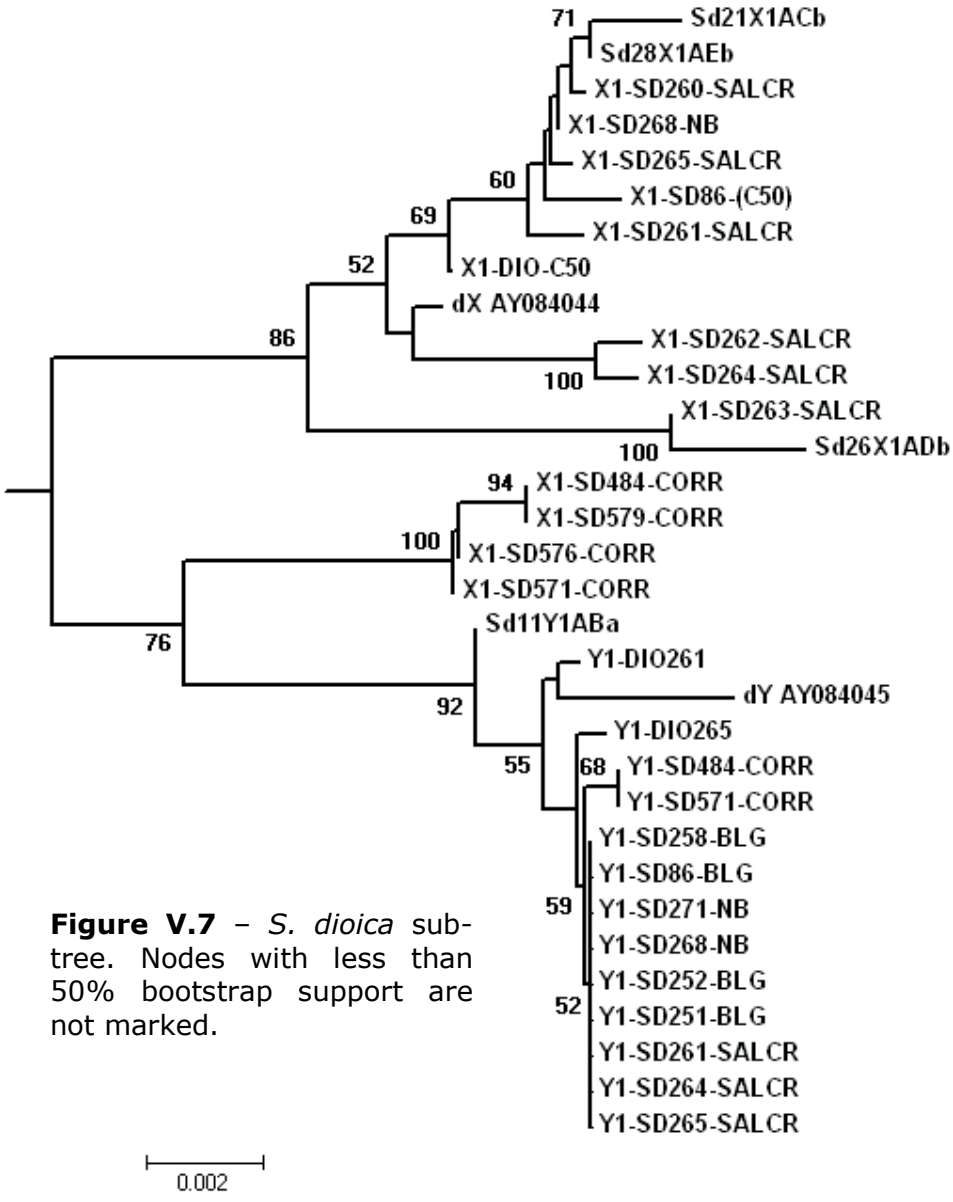


Figure V.7 – *S. dioica* sub-tree. Nodes with less than 50% bootstrap support are not marked.

V.3.3 Parsimoniously informative sites

The alignment below shows the parsimoniously informative sites taken from the 4.8kb alignment of Genbank sequences (the *S. dioica* sequences are labelled dX and dY). Of the 99 sites, 21 represent changes which occurred between the outgroup lineage and the sex chromosomal sequences. Interestingly there are 11 sites where a substitution between *S. vulgaris* and *S. conica* is still 'polymorphic' amongst the X and Y sequences (though it may be fixed in one of X or Y). Subsequently, there is little information to help elucidate the relationships of sequences within *S. latifolia* and *S. dioica*. Perhaps revealing, is that nine of the sites differing between *S. vulgaris* and *S. conica* are, at the same time, fixed differences between the Y chromosomes of *S.dioica* and *S. latifolia* and yet are polymorphic amongst the X chromosomes of these species. Many of the remaining sites illustrate the shared derived nature of the two *S. latifolia* Y sequences.

Parsimoniously informative sites from 4.8kb alignment of Genbank sequences

```
                                    111111 1111111111 1111111122 2222222222
    Site     111112233 3344588889 9999001112 2223333444 6667789923 3344677777
  Numbers   6002281313 6929503382 3688062330 2782578337 0341708870 2434411234
            9030694258 6302930913 4102962383 8924242783 5239060747 9142136004
VULGARIS    CCCGGCGC-- -----TTGTC TGTAATATTT TACACAGAAC ACCCAACCCT ACGGA-----
CONICA      ......A.AC TCGCG....T ......G... ......-... .....G.T.. ...A.-AGGG
X_AJ310656  .T..T..AAT TCGAA.AAGT .......CCC CTTGG..GTT .TG.G..TA. G...GAATGG
Y_AJ310655  TTGCTG.AGC GTTGGCAA.. GACGGCGCCC CT.GGGAGTT TTGTGCTT.C GTAA.TGTAA
Y_AY084037  TTGCTG.AGC GTTGGCAA.. GACGGCGCCC CT.GGGAGTT TTGTGCTT.C GTAA.TGTAA
X_AY084036  .T..C..AAC TTTGGCAAGT .....CGCCC CTTGG..GTT .TG.G..... G..AGAACGG
dX_AY084044 .T..T..AAT TCTAA-AA.. .....CGCCC CT.GG..GTT .TG.G.T.A. G..AGAACGG
dY_AY084045 .T..T.AAAC TCTAG.AA.. .A.G.CGCCC CT.GG..GTT .TG.G..... G..AGAATGG

            2222222333 3333333333 3344444444 444444444
            7788999112 2234555667 8833333334 566777888
            5907266133 4534022138 5811222573 518359256
            4818517498 4692849989 4138178946 588584699
VULGARIS    ---------T TGCTACGACA TACGATCTTC CGCGCA--C
CONICA      TTATGTCAA. .A...T...G C.-------- -C.T.TT--
X_AJ310656  TTGTTCTAA. GA.C..CG.G .....GT..T ..T...CC.
Y_AJ310655  TGGCCTCAGA G.TCCTCGTC CCAAG..GAT GCT.TGCCT
Y_AY084037  TGGCCTCAGA G.TCCTCGTC CCAAG.TGAT GCT.TGCCT
X_AY084036  CTATCCTAG. G..C..CGTG C.A.G.TGA. GC...GCG.
dX_AY084044 CTATCCTTG. G..C.TCGTG C.....TG.. ..T..GCG.
dY_AY084045 TTGTTCCTG. G..C.TCGT. C....GTG.. .C.TT.TG.
```

Gaps marked with '-' Identities marked with '.'

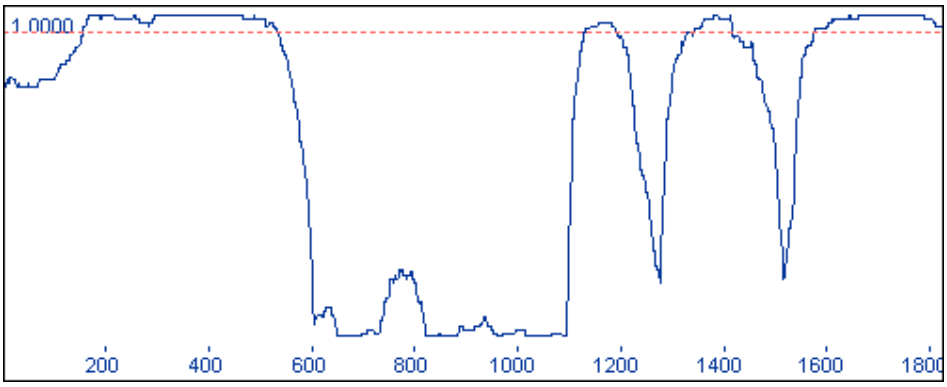
V.3.4 Recombining sequences

The four-gamete test was used to test if recombination had occurred between the X and Y chromosomal sequences. Using the 4kb Genbank sequences, with one X and one Y sequence from *S. latifolia* and the X and Y sequences from *S.dioica*, the minimum number of recombination events between the sequences was 6 or 11 depending on which pair of the *S. latifolia* sequences were used. However, when either X sequence was replaced with the unused *S. latifolia* Y sequence, the number of recombination events reduced to zero. The converse situation, with the three Y chromosomal sequences and one of the X sequences reduced the minimum number of recombinational events to zero. These tests confirmed the signal of recombination with SlXY1, but from these tests, it was not clear whether recombination was confined to X or Y or had occurred between the two.

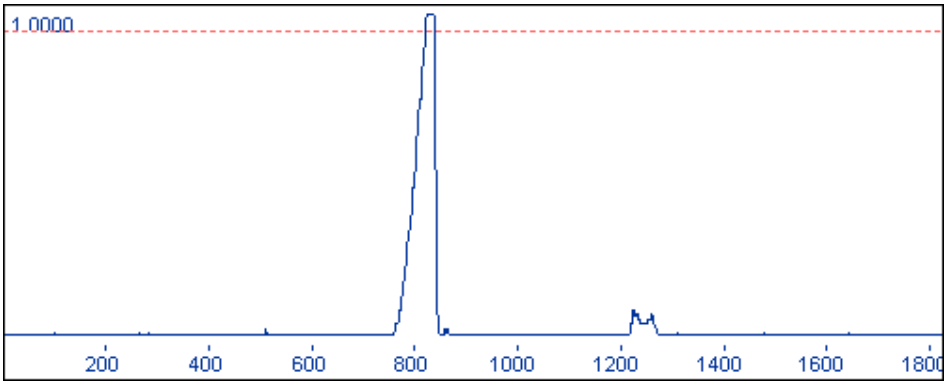
To cast the net wider, the same test was used on larger samples of X and Y sequences, this time confining the test to just X or Y sequences to search for the presence of recombination within one chromosome. The length of alignment used was 1.8kb. Using 38 Y sequences from *S. dioica* and *S. latifolia* (including the sequence from Sa14), no recombination events were detected by the four-gamete test. With 40 X sequences from *S. dioica* and *S. latifolia*, the minimum number of recombination events is 18. This would support the previous findings that there is recombination within the X sequences, but not within the Y sequences.

The sequence from Sa14, which appeared to fall between X and Y, was investigated to see if it had been involved in a recombination between X and Y. Firstly, this involved adding a single X sequence (X AJ310656) to the set of Y sequences and re-running the four-gamete test. When Sa14 was amongst the set of sequences, the minimum number of recombination events increased to 2, whereas without Sa14, the number remained at zero. This effect was not seen when the alternative X sequence (X AY084036) was used instead. This sequence was then investigated using the Hidden Markov Model implemented in TOPALi (described in Chapter IV) (Milne et al. 2004). The sequences SlX1_Sa14, X AJ310656 (*S. latifolia* X), Sd11Y1Aba (*S.dioica* Y) & AJ310655 (*S. latifolia* Y) were used to measure their support for the three possible topologies relating them. In the absence of recombination, one topology should be favoured along the length of the sequence. The graphical output of this analysis is shown in Figure V.8. There is a strong changeover of support between topologies 1 and 3 of Figure V.8 which group the Sa14 sequence with the *S. latifolia* X or *S. latifolia* Y sequences respectively. The support changes back from topology 3 to topology 1 before the end of the alignment, suggesting that just a short region in the centre of the Sa 14 sequence is homologous to the Y chromosomal sequence. The spike of support for topology 2 around 800bp into the alignment lies in a short region (<50bp) of micro-satellite repeats in which the alignment may well be unreliable.

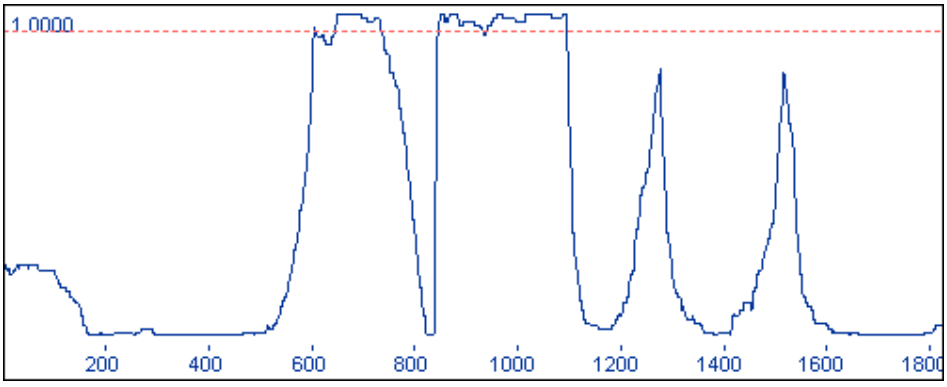
Figure V.8 – Graphical output of HMM run from TOPALi measuring support for three possible topologies between Sa14 sequence and X and Y from *S. latifolia*.



Topology 1: (SlX1_Sa14, X AJ310656), (Sd11Y1ABa, Y AJ310655)



Topology 2: (SlX1_Sa14, Sd11Y1ABa), (X AJ310656, Y AJ310655)



Topology 3: (SlX1_Sa14, Y AJ310655), (X AJ310656, Sd11Y1ABa)

Unfortunately, with the four sequences used in this analysis it is not possible to know if the clustering of the two sequences Sa14 and *S. latifolia* Y is because they are homologous or if in this region, the *S.*

latifolia X sequence bears closer homology to the *S. dioica* Y sequence. A more distant outgroup sequence would help, but the number of deletions in the available outgroup sequences relative to the *S. latifolia* and *S. dioica* sequences reduce the strength of signal to such an extent that it disappears.

Instead, we can look at the parsimoniously informative sites relating these sequences and the outgroups (shown below).

Parsimoniously informative site along the 1.8kb alignment

```
                                1111111111 1111111111 1111111111 1111111111 111
    Site          12222334 4555566666 0000000001 1112334444 4444555555 5566667777 777
   numbers   3660355462 7017805678 4444558890 4678341366 7899234578 9913590122 345
             1254516760 6870297742 1369567832 7584477824 2723351588 4839580308 653
SlX1_Sa14    CC-TATCGCT CCGGCCCGAT AAAGTTGCCA ATAGGCTTAG TAAAGTCCGC CCTCTGAGAC CTG
X_AJ310656   T.C..C.... ....TAATGG C.GAG.T..T ...C.G.... .C..AC.... .......... ...
Y_AJ310655   .TC.C...T. ..C....... .....C.... .......... .T...C.... .T.T.A..G. .G.
Y_AY084037   .TC.C...T. ..C...---- .......... .......... .T...C.... .T.T.A..G. .G.
X_AY084036   ..C..C..T. .....A---- ..G....... .CT...C... .C...C..-G ..--------- ---
dX_AY084044  ..C.....T. .....A---- C.GA.....T .C.C.G.... .C...C..-G ...-------- ---
dY_AY084045  ..C.....T. .TA..A---- C.GAG....T .C.C..C..T .T..AC.T-G ....-...G. ..A
VULGARIS     ..T..CGA.. T.AATA-TGG CGGA.CTTTT TC.CAGCCC. GCGGACG--- A.C.G.GCGG T..
CONICA       ..TG..GA.C -....AA--- ---------- ----T.C..T .C.TT..TA- ----....G. ..T
FLOS-JOVIS   T.TG.CTA.C TT.A.A---- CGGA...TTT TCTCA.CCC. GGGG.CG.A. A.C.G.GCGG T.A
```

Gaps marked with '-' Identities marked with '.'

The changeover in similarity of Sa14 to X or Y around 600bp into the alignment seems to be born out by the parsimoniously informative sites. However, the position of the change back, which seems to be at around 1100bp in Figure V.8, looks slightly later based on these sites. Throughout, the *S. dioica* Y sequence seems to be distinct.

However, of the five Y differences towards the end of the sequence which are not shared by Sa14 (T,T,A,G,G), four were derived in the Y lineage (on comparison with the outgroups). This pattern could also be consistent with the placement of Sa14 outside of the remaining Y sequences but not recombining with the X. The Y derived mutations could then have

happened since the split of Sa14 and the rest of the Y population. However, the blocked nature of the similarity of Sa14 with either X or Y sequences does seem quite suggestive of a degree of gene conversion between the two.

V.4 Discussion

A 2kb fragment of the SlXY1 gene was sequenced from fifteen of the panel of twenty-three Silene plants. One sequence from *S. dioica* Sd11 was actually SlY1 (*S. dioica* homologue), and another, from *S. latifolia* Sa14 appears to be part SlX1 and part SlY1.

The dataset is broadly suitable for the planned measurement of linkage disequilibrium, though the lack of sequence from some individuals will reduce the power of the test relative to what would have been otherwise. Further attempts to amplify SlX1 from the missing individuals should be made. As only three SlX1 sequences were obtained from *S. dioica* plants and one of these was a strongly *S. latifolia* haplotype in a hybrid plant, there will be no power to measure linkage disequilibrium within this species and the number of shared polymorphisms detected between the species will be reduced.

The nature of the sequence obtained from Sa14 should be further investigated. It's chimerical combination of SlX1 and SlY1 sequence points to recombination between the X and Y chromosomes of *S. latifolia*, a contradictory conclusion to the clustering of all other X and Y sequences and the apparent lack of recombination within the Y chromosomes. Though this sequence was not cloned, none of the direct sequencing reads displayed heterozygous peaks in the sequence and there was no overlap of sequence within the chromatograms, as would be expected if the PCR product contained both X and Y products, which differ by many indels. The

lack of identity with the Y chromosome in some regions but not others may be due to an increased rate of divergence of the Y sequences in this species (Filatov and Charlesworth 2002). However, it is not known how this variation might be organised into discrete blocks along a sequence and at present, recombination seems a better explanation. As only one sequence has been taken from this individual, it would be beneficial to find if this individual carried another sequence that looked more like an X or Y sequence. A segregation analysis using a plant bearing the same sequence would be needed to confirm its linkage.

The phylogenies relating the sequences obtained for this study and those sequenced by D. Filatov, better support the independent cessation of recombination in the *S. dioica* lineage at this locus. Unfortunately, the missing link in this chain of reasoning is that the *S. dioica* SlY1 sequence was originally defined by a perceived similarity to the SlY1 sequence of *S. latifolia*. This now seems to be erroneous, but still, the perfect split of each pair of *S. dioica* products into two halves of the tree, suggests that the diagnosis of Y linkage is correct. Ideally, a segregation analysis of *S. dioica* products would complete the definition of SlX1 and SlY1 in this species.

Before the analysis of linkage disequilibrium has been carried out, this study has revealed that the pattern of cessation of recombination between X and Y in these species is more complicated than previously thought. Recombination, to some extent continued between SlX1 and SlY1 in at least one species after the speciation of *Silene latifolia* and *Silene dioica*.

VI THESIS DISCUSSION & CONCLUDING REMARKS

The discussion will focus on two key assumptions made in this thesis, which may not have received sufficient attention and, if broken, may influence some of the conclusions drawn. These are firstly, that the processes of molecular evolution operating on a gene are equivalent in divergent species and secondly, that measures of substitution at 'silent' sites are a fair reflection of the neutral mutation rate.

Whilst the work in Chapter II has been published and become part of the fast expanding literature of genomic comparison studies, the discussion in that chapter and the ensuing paper did not take into account a recent key paper with potentially significant implications for this kind of study. Kumar and Subramanian (2002) challenged the assumption that the evolution of the same piece of sequence will proceed in a like manner in sister taxa. If orthologous sequences have been under the influence of identical evolutionary forces (e.g. mutation), then there should be no significant difference in their compositions (homogeneity). However, if there had been a change in the forces acting on one or both sequences (e.g. one sequence translocated to a region in which substitutions from AT -> GC occur at a greater rate), then one of the sequences would accumulate substitutions at a different rate or of a different type relative to its homologue. Whilst with only two sequences it is impossible to know in which lineage a particular mutation/substitution occurred, Kumar and Gadagkar (2001) presented a test for heterogeneity based on the

difference in composition of two sequences. Aligned homologues failing this test are considered to be evolving in an "atypical manner" (Kumar and Gadagkar 2001). When orthologues from human and mouse – the dataset often used to test for patterns in the rate of evolution around genomes – were checked for disparity, over 40% of genes failed the test (Kumar and Gadagkar 2001; Kumar and Subramanian 2002). Between mouse and rat, the proportion was around 10%, apparently reflecting greater homogeneity of evolutionary rate within the rodents. Indeed, the authors suggest that the difference in degree of heterogeneity between different comparisons (e.g. mouse-human vs. mouse-rat) could have been a cause of previous findings of heterogeneity in the evolutionary rate within a genome. From the simulations of Kumar and Gadagkar (2001), the power of the disparity index test is strongly linked to the level of divergence between sequences (irrespective of natural causes like chromosomal rearrangements). Using around 200 sites (the average in Kumar and Subramanian 2002), the simulated power of the test is less than 20% for sequences with less than 10% divergence (Kumar and Gadagkar 2001). In addition, Lercher, Chamary and Hurst (2004) suggested that by only accepting adjacent genes syntenic in both human and mouse (Kumar and Subramanian 2002), the sample sizes in some of their groupings (over 1.5 Mb) became very small and prone to greater error variance. Lercher, Chamary and Hurst (2004) went on to rediscover genomic heterogeneity in the rate of synonymous substitution after accounting for disparate sequences by using an averaged level of local similarity which maintained a larger sample size. Whilst the lack of difference between near and far neighbours may still be argued, the small

sample size and method of presentation of variation amongst chromosomes in the paper by Kumar and Subramanian (2002) was not persuasive.

Although the disparity index was promoted for choosing a subset of genes on which to base a molecular phylogeny, other implications quickly became apparent. After removing heterogeneously evolving orthologous gene pairs, Kumar and Subramanian (2002) found no heterogeneity in the rate of silent site evolution between genes on the same chromosome. Furthermore, they found no significant difference between orders of mammals for the estimates of the base mutation rate (inferred from the neutral substitution rate), which would suggest a global molecular clock for mammals and no generation time effect (Wu and Li 1985). If this were true, then replication-independent processes would account for more mutations than those associated with replication itself.

Yi, Ellsworth and Li (2002) used genes for which sequence from more than two species was available to measure the relative rates of evolution in different lineages. Using New World monkeys as an outgroup, they concluded that the rate of substitution in old world monkeys was greater (by ~30%) than in apes. Surprisingly, this effect was slightly stronger in orthologous genes for which the above disparity test did not find any heterogeneity in the pattern of substitution. The inference from this study was that a generation time effect may indeed exist and that the rate of evolution in apes has slowed relative to that of our immediate ancestors.

Of immediate interest to this thesis is whether the disparity index test would have had an effect on the results presented in Chapter II? Whilst the hypothesis under examination concerned a difference in the rate of substitution between the PAR and the rest of the genome, the inferences on the relative rates of mutation depend on an equivocal relationship between mutation and substitution in both species. Using the test of Kumar and Gadagkar (2001), no disparity was detected in any of the gene sequences from the PAR1, which is conserved between humans and the other apes, nor in SYBL, the only gene from the human-specific PAR2. Of the 51 intronic sequences which made up the genomic distribution, not one failed the disparity index test using the recommended cut-off value of $P<0.01$ (Kumar and Gadagkar 2001). However, five of them showed disparity using $P<0.05$, slightly more than one would expect from Type I errors under the null hypothesis of homogeneity. The proportion of disparate genes is similar to those between mouse and rat, and although the power of the test is low in close comparisons (Kumar and Gadagkar 2001), it should reflect general homogeneity of evolutionary patterns between the human and orang-utan genomes. Either way, the removal of five sequences does not alter the distribution of divergences and, as the divergences of the PAR sequences do not overlap, it does not alter the conclusion that the substitution rate is much greater in the PAR. The negative result from Chapter II, concerning the lack of correlation of recombination rate with the 51 intronic sequences, is similarly unaffected by the removal of these five sequences.

The use of the substitution rate in Chapter II, also assumes neutrality of the intronic mutations. The work presented in chapters III & IV was carried out under the assumption that Ks, the synonymous rate of substitution, is a also fair measure of the neutral rate of evolution and unaffected by selection (Kimura 1983). The measurement of the strength and mode of selection operating on the genes studied in these chapters relies on a qualitative difference in the ability for that selection to operate on non-synonymous and synonymous changes. Positive selection is inferred when the rate of fixation of amino acid changing mutations appears greater than that for 'silent' changes.

Whilst it was shown some time ago that the bias in the use of certain synonymous codons over others is related to gene expression level for bacteria (Sharp and Li 1986), yeast (Bennetzen and Hall 1982) and Drosophila (Shields et al. 1988), until very recently there did not seem to be a clear association in mammals. This was generally assumed to reflect a reduction in the efficacy of selection in mammals relative to invertebrates due to the substantially lower population size of mammals (Li 1987). Where there was bias within a gene, it was linked to mutational forces such as the local GC composition rather than selection for optimal codons (Sharp et al. 1995). However, several independent lines of research have recently contradicted the assumption that synonymous sites are evolving neutrally in mammals.

Iida and Akashi (2000) found that within the exons of individual human genes, the exons that were sometimes spliced from the mRNA had lower

codon bias than the exons that were always translated. This study suggested that selection was operating to maximise the translational efficiency of mRNAs.

Following on from an observation that highly expressed mammalian genes tend to have shorter transcripts (Castillo-Davis et al. 2002), Urrutia and Hurst (2003) were able to show that the total intron length was reduced in these genes while the number of introns was not, suggesting that there was a selective advantage to possession of introns not associated with gene length (e.g. reduction of Hill-Robertson interference) (see also Comeron 2004). A structural role for introns at the mRNA stage would be consistent with this hypothesis and as the structure and stability of RNA is specified by its nucleotide sequence and composition, these would come under selection. Furthermore, after correcting for base composition and gene length Urrutia and Hurst (2003) did find a significant correlation between codon bias and expression level in human genes. This seemed to be mostly confined to the 30% or so of most highly expressed genes in their study. The effect is there but it is significantly weaker and therefore harder to find than in Yeast or Drosophila. So, it is still consistent with a reduced efficacy of selection.

Interestingly, Plotkin, Robins and Levine (2004) found organisation of human codon bias at another level, namely that of tissue specificity. They found that tens of genes expressed in the same tissues (e.g. brain) were alike in codon bias and that in many cases this same bias had been preserved in the same genes in mouse (Plotkin, Robins, and Levine 2004).

Genes from different tissues (e.g. testis vs. uterus) were biased towards different sets of codons. The effect was strong enough that the tissue specificity of many genes could be deduced from their codon bias. The authors predict that tRNA abundance may vary with tissue type in mammals. It also has yet to be seen whether this trend holds for all tissue specific genes or whether there is also an effect of expression level.

Bustamante, Nielsen, and Hartl (2002) measured the rates of synonymous and non-synonymous substitutions in pseudogenes relative to their existing functional genes. Whilst they found a strong effect of amelioration of pseudogenes (both silent and non-silent sites accumulated changes to fit into the regional composition) (see also Subramanian and Kumar 2003), they also found that in those genes that had moved to a region with similar composition, the silent sites picked up slightly more substitutions than their counterparts in the surviving functional genes, suggesting relaxation of functional constraint on these sites.

Seffens and Digby (1999) investigated the structural stability of natural mRNAs by calculating *in silico* their folding energies. They found that, after controlling for the lengths and GC% of the mRNAs, and keeping the amino acid content the same, using different codons adversely affected the stability of the mRNA molecules. Shen et al. (1999) used structure specific nuclease enzymes to show that naturally occurring single nucleotide polymorphisms (SNPs) could change the structure of the transcribed mRNA. There is no reason why such SNPs wouldn't be located in otherwise synonymous sites, and recently a 'silent' single nucleotide

polymorphism in human dopamine receptor D2 was shown to alter the structure of the mRNA enough to affect the level of translation deleteriously (Duan et al. 2003). The occurrence of this polymorphism in the human population has been linked to several psychological disorders (Duan et al. 2003).

Whilst it is becoming clear that codon bias is operating in some mammalian genes and that synonymous sites and introns are not as neutral as previously thought, the difficulty with which these effects were found and the ease with which they are overpowered by other forces (such as mutation bias), suggests that selection for synonymous changes in mammals operates at a significant level only in the most highly expressed genes. For most genes, selection on synonymous sites is still relatively weak compared to that for non-synonymous changes.

The results of Chapter II are unchanged if disparity is taken into account because the distributions of divergence from the PAR genes and the rest of the genome do not overlap. There is also no reason to suspect that the introns of all genes are under strong purifying selection except those in the PAR. A difference in the base mutation rate is a more acceptable explanation. For the comparisons of synonymous versus non-synonymous sites in Chapters III and IV, selection operating to conserve the synonymous sites would inflate the apparent difference at putatively positively selected non-synonymous sites. However, the order of magnitude difference seen for most genes in the rate of substitution at these sites clearly illustrates that the average amino-acid replacing

mutation is heavily selected against (Mouchiroud, Gautier, and Bernardi 1995; Makalowski and Boguski 1998).

Concluding Remarks

The sex chromosomal studies presented in this thesis have measured the rate of molecular evolution in regions with different levels of recombination. Firstly, the contraction of the mammalian pseudoautosomal region to such a small region has created the extreme rate of recombination used as a contrast to the rest of the genome in Chapter II. From the increase in substitution rate relative to the recombination rate we found in the PAR, it could have been predicted that this effect would be fairly weak around the genome. Since this was done in 2002, several other studies using more data from around the genome have supported this finding. However, the 'mutagenic effect of recombination' remains a correlation and the next priority is to find a mechanism to relate the two.

In Chapters III and IV, the investigation of evolutionary models using maximum likelihood estimation showed contrasting rates of substitution between primate X and Y sequences. Even though the forces of degeneration reduce the efficacy of selection on the Y chromosome, the effect is no longer so great that genes cannot survive and even adapt. This may depend on the number of genes segregating on the Y, which would affect the deleterious mutation rate. Intuitively, it seems reasonable that, with a constant mutation rate (per locus) and constant

population size, the long-term genetic content of a Y chromosome may level out.

Whether AMELY will survive the thinning out process is difficult to tell. Its expression profile does not match that of the hardy genes SMCY, UTY and USP9Y, but it appears to be still functional on all the Y chromosomes sequenced and maintains conserved sequence whilst many other stratum-four genes have become pseudogenes in the human lineage alone, suggesting that it has not yet given up.

Whilst it may be that the rate of degeneration of the mammalian Y chromosome has slowed down, it has barely started in *Silene*. However, the restriction of recombination, which does seem to have acted in a stepwise manner akin to the mammalian sex chromosomes, has already taken over a significant percentage of the Y chromosome. Chapter V revealed the independent cessation of recombination between SlX1 and SlY1 in both *S. dioica* and *S. latifolia* and suggests that both species inherited a directional force already operating in their immediate ancestor.

The inexorable restriction of recombination on the *Silene* sex chromosomes, and expected degeneration that will follow, paints a bleak future for the non-recombining chromosome in these plants. More reassuring though, is that this process may have a limit, and that my Y chromosome, being an experienced Y chromosome, will still be around in years to come.

VII APPENDIX

Chapter IV Amelogenin sequence alignment A2 in NEXUS format.

Exon positions

Exon 5: 1 - 56

Intron 5: 57 - 347

Exon 6: 348 - 840

Sequence names and species.

Sequence names are prefixed with a species identifier and suffixed with X or Y. Cow (Bos), Goat (Capra), Japanese serow (Jcap), Pig (Sus), Human (Hsap), Chimpanzee (Pan), Golden hamster (Mes), Mouse (Mus), Rat (Rat), Squirrel monkey (Saim), Small eared galago (Otole), Ring-tailed lemur (Lemur), Horse (Equus).

A09 ANb & A09 AMc are distinct sequences from spider monkey

M10 AOa & M10 APb are distinct sequences from macaque

```
#NEXUS
[ Title : file generated by PROSEQ v 2.9 pre-beta]
begin data;   dimensions ntax=24 nchar=840;
   format missing=? gap=- matchchar=. datatype=nucleotide interleave;
   matrix

[!Domain=Data;]
[                   1 1111111112 2222222223 3333333334 4444444445 5555555556 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  tttctcttaa ggtgcttacc cctttgaagt ggtaccagag c---ataagg ccaccggtat
PanAMELX   .......... .......... .......... .......... .---...... ..........
G73AMELX   .......... ......c... .......... .......... .---...... ..........
M10_AOa    .......... .......... .......... .......... .---...... ..........
P70AMELX   .......... ......c... .......... .......... .---...... ..........
C71AMELX   ---------- --------.. .......... .......... .---...... ..........
SaimAMELX  .......... .......... .......... .......a.. .atg...... ...g.....c
T65AMELX   .......... ......c... .......... .......a.. .atg...... ...g.....c
A09_ANb    .......... ...c...... .......... .......a.. .atg...... ...g.....c
A09_AMc    .......... ......c... .......... .......... .attg..... .....a...c
SaimAMELY  .......... ......c... .......... .........c .atg...... ...t.a....
M10_APb    .......... ......c... .......... .......... .atgg....a .....a...c
C71AMELY   ------.... ......c... .........c .......... .atg.....a .....a...c
G72AMELY   .......... ......c... .......... .......... .atg.....a .....a...c
P11AMELY   .......... ......c... .......... .......... .atg.....a .....a....
G73AMELY   .......... ......c... .......... .......... .atg.....a .....a....
HsapAMELY  .......... ......c... .......... .......... .atg.....a .....a....
PanAMELY   .......... ......c... .......... .......... .atg.....a .....a....
LemurAMELX .....t.... .......... .......... .......... .atg...... .........g
OtoleAMELX .......a.. .......... ...g...... .......... .atgt....a .........g
MusAMELX   .......... .......... .......... .......... .atg...... .ag.......
EquusAMELX .......... .......... ...c...... .......... .ttg..c... .ag.......
EquusAMELY .......... .......... ...c...... .........a .atg.....a .ag.......
AllExons   ---------- -......... .......... .......... .---...... ......----

[                                                     1 1111111111 1111111111 ]
[          6666666667 7777777778 8888888889 9999999990 0000000001 1111111112 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  gtagacattt tgttccttat tccctgaaaa tattaggcat gcattaaaat tcccatatta
PanAMELX   .......... .....t.... .......... .......... .......... ..........
G73AMELX   .......... .....t.c.. .......... .......... .......... ..........
M10_AOa    .....t.... .....t.... .......... .......... .......... ..........
P70AMELX   .....t...c .....t.... .......... .......... .......... ..........
C71AMELX   .....t.... .....t.... .......... .......... .......... ..........
SaimAMELX  .....t.... cc...t.... .......... .......... .....t.... ..........
T65AMELX   .....t.... .c...t.... .......... .......... .....t.... ..........
A09_ANb    .....t.... .c...t.... ......c... .......... .....g.... ..........
A09_AMc    .....t.... .....t.... ...a....t. ..a....... .....t.... ......t...
SaimAMELY  ...c.tg... .....t.... ...a.t..t. ..a....... .....t.... ....g.t...
M10_APb    .......... .....t---- ..t......t .......... .....t.... ......t..c
C71AMELY   .......... ...c...... .......... .......... .....t...c ......t...
G72AMELY   .......... .....t.... .......... .......... .....t...c ......t...
P11AMELY   .......... .....t.... .......... .......... .....t.... ...t..t...
G73AMELY   .......... .....t.... .......... .......... .....t.... ......t...
HsapAMELY  .......... .....t.... .......... .......... .....t.... ......t...
PanAMELY   .......... .....t.... .......... .......... .....t.... ......t...
LemurAMELX ...----... .....t.... .....t.... ......-... .....t.... ..........
OtoleAMELX ...----... .....t.... .....t.... .......... .....t.... ..........
MusAMELX   ...a...... .....t.cc. .....t..g. g.a...a... .tg..tt..g ....t.t...
EquusAMELX .....-.... .-...t.... .....t.... ...a.at... .....tc... ....-.t...
EquusAMELY .....-.... .....t.... .....t.... .....a.t.. .g...tc... ...t..t...
AllExons   ---------- ---------- ---------- ---------- ---------- ----------
```

```
[          1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 ]
[          2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  agtgaaatat catgtctact ccacatgcag acattaatgg gaaatttagt ttgtaaaaa-
PanAMELX   .......... .......... .......... .......... .......... .........-
G73AMELX   .......... ........g. ........g. .......... .......... .........-
M10_AOa    .......... .......... ......a... .......... .......... .........-
P70AMELX   .......... .......... ......a... .......... .......... .........-
C71AMELX   .......... .......... ......a... .......... .......... .........-
SaimAMELX  .......... ---..t.... ......a.t. g..c...ga. .......... .........a
T65AMELX   .......... ---..... ......a.t. ...c....a. .......... .........a
A09_ANb    .......... ---..... ......a.t. ...c....c. .......... .........a
A09_AMc    ..-a...... .....g.... ......a... ...c...... .......... .........-
SaimAMELY  ..-a...... ...a.a.... t.....a... ...c...... .......... .........-
M10_APb    ..-a...c.c t...cg.... ......a... ...c...... .......... .........-
C71AMELY   ..-a.....c .....g.... ......--.. ...c...... .......... .........-
G72AMELY   ..-a.....c .....g.... ......--.. ...c...... .......... c........-
P11AMELY   ..-a.....c .....g.... ......a... ...c...... .......... .........-
G73AMELY   ..-a.....c .....g.... ......a... ...c...... .......... .........-
HsapAMELY  ..-a.....c .....g.... ......a... ...c...... .......... .........-
PanAMELY   ..-a.....c .....g.... ......a... ...c...... .......... .........-
LemurAMELX .......... ---..... ......ag.. ..cc...... ...g...... .-......g-
OtoleAMELX ........g. ---..... ......a... ...c...... ..t..-.... .-.....gg-
MusAMELX   .c........ ...a...... ....c.t..a ...c...... ...t...... ........g-
EquusAMELX ...a.....a ...a....tg ......a... ...c.....c .....g..a. ........g-
EquusAMELY .........a ...a.t..tg ......a... ...-..t... .....c.... .........-
AllExons   ---------- ---------- ---------- ---------- ---------- ----------

[          1111111111 1111111112 2222222222 2222222222 2222222222 2222222222 ]
[          8888888889 9999999990 0000000001 1111111112 2222222223 3333333334 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  ---atcatat ctgtgtacac agttacaaa- tttttgca-a aggaaaaatg aataaa----
PanAMELX   ---....... .......... .........- ........-. .......... ......----
G73AMELX   ---....... .......... .........- ........-. .......... ......----
M10_AOa    ---....... .......... .........- ........-. .......... ......----
P70AMELX   ---....... .........t .........- ........-. .......... ......----
C71AMELX   ---...g... .......... .........- ........-. .......... ......----
SaimAMELX  aaa....... ..t......t .........- ........-. .......... ......----
T65AMELX   ---....... ....t....t .....a...- ........-. .......... ......----
A09_ANb    ---c...... .......... .....t...- ........-. .......t.. ......----
A09_AMc    ---c..c... .a.......t g......c.- ......g.-. .......... ......----
SaimAMELY  ---c..c... .at....... .......c.- ......g.-. .......... ......----
M10_APb    ---....... ..-....... .........- ........-g ......---- ......----
C71AMELY   ---g...... .......... .........- ........-. .........a ......----
G72AMELY   ---g...... ....t..... .........- ........-. .........a ......----
P11AMELY   ---....... .......... .c.......a ........-. .........a ......----
G73AMELY   ---....... .......... .........- a.......-. ......g..a ......----
HsapAMELY  ---.....g. .......... .........- ........-. ......g..a ....c.----
PanAMELY   ---.....g. .......... .........- ........-. ......g..a ......----
LemurAMELX ---g....g. g......... .....g...- ..c.....g. ..a.g..... ......----
OtoleAMELX ----...... g......... .....g...- ..c.....g. ..a....... ......----
MusAMELX   ---g...... .......... t....g...- ..c.....-. .......... .c....----
EquusAMELX ---c...... .........t .....g...- ..c.....-c .......... ......----
EquusAMELY ---g....g. .........t .a...g...- .a...t.c-c ..a..g.... ......catc
AllExons   ---------- ---------- ---------- ---------- ---------- ----------
```

```
[          2222222222 2222222222 2222222222 2222222222 2222222222 2222222223 ]
[          4444444445 5555555556 6666666667 7777777778 8888888889 9999999990 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  -atattccta tagccata-- -------atg gcaaagaaaa cactgctgct tctctggttg
PanAMELX   -......... ........-- -------... .......... .......... ..........
G73AMELX   -......... ........-- -------... .......... .......... ..........
M10_AOa    -...c..... ........-- -------... .......... .......... ..........
P70AMELX   -......... ........-- -------... .......... .......... ..........
C71AMELX   -......... ........-- -------... .......... .......... ..........
SaimAMELX  -....c.... .g......-- -------... at........ .......... ..........
T65AMELX   -......... .g......-- -------... .t........ .......... ..........
A09_ANb    -......... .g......-- -------... .t........ .......... ..........
A09_AMc    -......... .g......-- -------... .t........ .......... .......c..
SaimAMELY  -......... .gaa...t-- ------ag.. .......... .........a. .......c..
M10_APb    -........g gg......-- -------... .......... .......... ..........
C71AMELY   -......... .g......-- -------... ...g...... .......... .......g..
G72AMELY   -......... .g......-- -------... ...g...... .......... .......g..
P11AMELY   -........g .g....c.-- -------... .......... ......c... .......a..
G73AMELY   -........t .g......-- -------... .......... .......... ..........
HsapAMELY  -......... .g......-- -------... .......c.. .......... ..........
PanAMELY   -......... .g......-- -------... .......... .......... ..........
LemurAMELX -......... .g......-- --------.. ......c... .......... .g..ca....
OtoleAMELX a......... .g......-- -------... ......c... .......... .g..ca....
MusAMELX   -......a.. ag.t....-- -------..a a....c.ct. .--....a.. ....ca...a
EquusAMELX --.....a.. .t.t....-- -------..a a......... ..t....a.. ....ca....
EquusAMELY c.....at.. ..atg...ta ataatga..c a......... ..t....ata ....ca.ca.
AllExons   ---------- ---------- ---------- ---------- ---------- ----------

[          3333333333 3333333333 3333333333 3333333333 3333333333 3333333333 ]
[          0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  g-agtcacct gagccaatgg taaacctgcc tctctgtttc tcaccagtac ccttcctatg
PanAMELX   .-........ .......... .......... .......... .......... ..........
G73AMELX   .-........ .......... .......... .......... .......... ..........
M10_AOa    .-........ .......... .......... .......... .......... ..........
P70AMELX   .-........ .......... .......... .......... .......... ..........
C71AMELX   .-........ .......... .......... .......... .......... ..........
SaimAMELX  .-........ .......... .......... .......... .......... ..........
T65AMELX   .-........ .......... .......... .......... .......... ..........
A09_ANb    .-........ .......... .......... .......... .......... ..........
A09_AMc    c-.......c .......... .......... .......... ...t...... ..c.......
SaimAMELY  c-.....g.. .......... .......... .......... ....t..... ..........
M10_APb    .-........ .......... .......... .......... .......... ..........
C71AMELY   .-......t. .......... .......... .......... c......... t.........
G72AMELY   .-......t. .......... .......... .......... c......... t.........
P11AMELY   .-.c...... ........a. .......... .......... .......... t.........
G73AMELY   .-........ .......... .........g .......... .......... t.........
HsapAMELY  .-......g. .......... .........a .......... .......... t.........
PanAMELY   .-........ .......... .........g .......... .......... t.........
LemurAMELX .a...g.... .....----- ---------- .......... .......... ..........
OtoleAMELX .a...g.... .....----- ---------- .......... .......... ..........
MusAMELX   .a.a...a.. .......... ........a. ...t...... .........t ..........
EquusAMELX .a.....atg .......... ........a. ...t...... .......... a.........
EquusAMELY ag.....atg .......... ........a. ...t.t.... .......... ..........
AllExons   ---------- ---------- ---------- ---------- -------... ..........
```

```
[          3333333333 3333333333 3333333333 3333333334 4444444444 4444444444 ]
[          6666666667 7777777778 8888888889 9999999990 0000000001 1111111112 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  gttacgagcc catgggtgga tggctgcacc accaaatcat ccccgtgctg tcccaacagc
PanAMELX   .......... .......... .......... .......... .......... ..........
G73AMELX   .......... .......... .......... .......... .......... ..........
M10_AOa    .......... .......... .......... .......... .......... ..........
P70AMELX   .......... .......... .......... .......... .......... ..........
C71AMELX   .......... .......... .......... .......... .......... ..........
SaimAMELX  .......... .......... .......... .......... ...t...... ..........
T65AMELX   .......... .......... .......... .......... ...t...... ..........
A09_ANb    .......... .......... .......... .......... .......... ..........
A09_AMc    .......... .......... .......... .......... .......... ..........
SaimAMELY  .......... t......... ......t... .......... ...t...... ..........
M10_APb    .......... .......... .......... .......... .......... ..........
C71AMELY   ......g... .......... .......... .......... ....a..... ..........
G72AMELY   ......g... .......... .......... .......... ....a..... ..........
P11AMELY   .......... .......... .......... .......... ....a..g.. ..........
G73AMELY   .......... .......... .......... .......... .......g.. ..........
HsapAMELY  .......... .......... .......... .......... .......g.. ..........
PanAMELY   .......... .......... .......... .......... .......g.. ..........
LemurAMELX .......... .......... .......... .......... .......... .....g....
OtoleAMELX .......... .......... .......... .......... .......... ..........
MusAMELX   .......a.. .......... .......... .......... ...t...... ..t.......
EquusAMELX .......a.. .......... .......... .......... .......... .....g...a
EquusAMELY ....t..a.. ....a..... .......... .......... ...t...... a....g...a
AllExons   .......... .......... .......... .......... .......... ..........

[          4444444444 4444444444 4444444444 4444444444 4444444444 4444444444 ]
[          2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  acccccccgac tcacaccctg cagcctcatc accacatccc agtggtgcca gctcagcagc
PanAMELX   .......... .......... .......... .......... .......... ..........
G73AMELX   .......... .......... .......... .......... .......... ..........
M10_AOa    .......... .......... .......... .......... .......... ..........
P70AMELX   .......... .......... .......... .......... .......... ..........
C71AMELX   .......... .......... .......... .......... ...t...... ..........
SaimAMELX  .......... .......... .......... .......... .......... ..........
T65AMELX   .......... .......... .......... .......... .......... ..........
A09_ANb    .......... .......... .......... .......... .......... ..........
A09_AMc    ....a..... .......... .......... .......... .......... ..........
SaimAMELY  .......... .......... .......... .......... .......... ..........
M10_APb    .......... .......... .......... .......... .......... ..........
C71AMELY   ....a.t... .......... .........- --........ .......... ..........
G72AMELY   ....a.t... .......... .........- --........ .......t.. ..........
P11AMELY   ......t... .......... .......... .........g .......... ..........
G73AMELY   ......t... ...t...... ...t...... .......t.. .......... ..........
HsapAMELY  ......t... .......... ...t...... .......... .......... ..........
PanAMELY   ......t... ...t...... ...t...... .......... .......... ..........
LemurAMELX ......a... .......... .......... .......... ......a.a. ..........
OtoleAMELX .......... .......... .......... .......... .......... ........a.
MusAMELX   .t.......g .........t .......... .....c.t.. c......... .....a....
EquusAMELX .t...t.a.a ....g..t.. .......... ....t..... ca.....t.g ........c.
EquusAMELY .t...t.a.a ....g..... .......... .......... ca........ ..c.......
AllExons   .......... .......... .......... .......... .......... ..........
```

```
[          4444444444 4444444445 5555555555 5555555555 5555555555 5555555555 ]
[          8888888889 9999999990 0000000001 1111111112 2222222223 3333333334 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  ccgtgatccc ccagcaacca atgatgcccg ttcctggcca acactccatg actccaatcc
PanAMELX   .......... .......... .......... .......... .......... ..........
G73AMELX   .......... .......... .......... .......... .......... ..........
M10_AOa    .....c.... .......... .......... .......... .......... .......c..
P70AMELX   .....g.... .......... .......... .......... .......... .......c..
C71AMELX   .....c.... .......... .......... .......... ........c. .......c..
SaimAMELX  .....g.... .......... .......... .......... .......... .t.....c..
T65AMELX   .....g.... .......... .......... .......... .......... .t.....c..
A09_ANb    .....g.... .......... .......... .......... .......... .t.....c..
A09_AMc    .....g.... .......... ........t. .......... .......... .t.....c..
SaimAMELY  .t...g.... .......... .......... .......... .......... .t....gc..
M10_APb    .....c.... .......... .......... .......... .......... .......c..
C71AMELY   ..ag.g...g .......g.. c.......t. .......... g..a...... .......c..
G72AMELY   ..cg..c... .......... c.......t. .......... g..a...... .......c..
P11AMELY   ...g.g.... a..a...... c.......t. .......... g..a...... .......c..
G73AMELY   ..ag.g.... .......g.. c.......t. .......... g..a...... .......c..
HsapAMELY  ..ag.g...g .......g.. c.......t. .......... g..a...... .......c..
PanAMELY   ..ag.g.... .......g.. ........t. .......... g..a...... .......c..
LemurAMELX .....g.... .......... .......... .......... .......... .......c..
OtoleAMELX .....g.... .......... .......... .......... .......... .......c..
MusAMELX   .....gc... .......... ........a. .......... c......... .......c..
EquusAMELX .t...g.... .......... ........ac .......... .......... gt.....c..
EquusAMELY .t...g..t. .......... gca..a..ac .......... .......... .t.....gt.
AllExons   .......... .......... .......... .......... .......... ..........

[          5555555555 5555555555 5555555555 5555555555 5555555555 5555555556 ]
[          4444444445 5555555556 6666666667 7777777778 8888888889 9999999990 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  aacaccacca gccaaacctc cctccg---- --------cc cgcccagcag ccctaccagc
PanAMELX   .......... .......... ......---- --------.. .......... ..........
G73AMELX   .......... .......... ......---- --------.. .......... ..........
M10_AOa    .......... .......... ......---- --------.. .......... ..........
P70AMELX   .......... .......... ......---- --------.. .......... ..........
C71AMELX   .......... ........c. ......---- --------.. .......... ..........
SaimAMELX  .......t.. .......... ......---- --------.. .......... ....t.....
T65AMELX   .......... .......... ......---- --------.. .......... ....t.....
A09_ANb    .......... .......... ....t.---- --------.. .......... ....t.....
A09_AMc    .......... .......... .....a---- --------.. .......... ....t.....
SaimAMELY  .......... ......tt.. .....a---- --------.a t.t....... ....t.....
M10_APb    .......... .......... ......---- --------.. .......... ..........
C71AMELY   .......t.. .......... ....t.---- --------.. t......... ....c.....
G72AMELY   .......t.. .......... ....t.---- --------.. t......... ....t.....
P11AMELY   .......t.. .......... ....t.---- --------.. t......... ....t.....
G73AMELY   .......t.g .......... ....t.---- --------.. t......... ....t.....
HsapAMELY  .......t.. .......... ....t.---- --------.. t......... ....t.....
PanAMELY   .......t.. .......... ....t.---- --------.. t......... ....t.....
LemurAMELX .t........ .......... ......---- --------.. .......... ....t.....
OtoleAMELX .......... .......... ......---- --------.. .......... ....t.....
MusAMELX   .......t.. .......a.. .....atccg cccagcag.. .tt....... ....t.....
EquusAMELX .......... .......... ......---- --------.. t.t....... ....t...c.
EquusAMELY .......... .......... .....a---- --------.. t.t....... ....t....-
AllExons   .......... .......... ......---- --------.. .......... ..........
```

```
[          6666666666 6666666666 6666666666 6666666666 6666666666 6666666666 ]
[          0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  cccagcctgt tcagccacag cctcaccagc ccatgcagcc c--------- ----------
PanAMELX   .......... .......... .......... .......... .--------- ----------
G73AMELX   .......... .......... .......... .......... .--------- ----------
M10_AOa    .......... .......... .......... .......... .--------- ----------
P70AMELX   .......... .......... .......... .......... .--------- ----------
C71AMELX   .......... .......... .......... .......... .--------- ----------
SaimAMELX  .....a.... .......... .......... .......... .--------- ----------
T65AMELX   .....a.... .......... .......... .......... .--------- ----------
A09_ANb    .....a.... .......... .......... .......... .--------- ----------
A09_AMc    .....a.... .......... .t........ .......... .--------- ----------
SaimAMELY  .t...a.... .......... t.......a. .......... .--------- ----------
M10_APb    .......... .......... .......... .......... .--------- ----------
C71AMELY   .......... .......... .......... .......... .--------- ----------
G72AMELY   .......... .......... ......a... .......... .--------- ----------
P11AMELY   .......... .......... .......... .......... .--------- ----------
G73AMELY   .......... .......... .......... .......... .--------- ----------
HsapAMELY  .......... .......... .......... .......... .--------- ----------
PanAMELY   .......... ........t. .......... .......... .--------- ----------
LemurAMELX .......c.. ......g... .......... .......... .atgcagccc atgcagccca
OtoleAMELX .......c.. ......g... .......... .......... .atgcagccc atgcagccca
MusAMELX   .....g.ca. ..ca..c... t....t.... .......... .--------- ----------
EquusAMELX .......... c.....g... .......... ....c..... t--------- ----------
EquusAMELY ---------- ----...... .......... ....c..a.. t--------- ----------
AllExons   .......... .......... .......... .......... .--------- ----------

[          6666666666 6666666666 6666666666 6666666667 7777777777 7777777777 ]
[          6666666667 7777777778 8888888889 9999999990 0000000001 1111111112 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  ---------- ---------- ---------- -----cagcc acctgtgcac cccatgcagc
PanAMELX   ---------- ---------- ---------- -----..... .......... ..........
G73AMELX   ---------- ---------- ---------- -----..... .......... ..........
M10_AOa    ---------- ---------- ---------- -----..... .......... ..........
P70AMELX   ---------- ---------- ---------- -----..... .......... ..........
C71AMELX   ---------- ---------- ---------- -----..... .......... ..........
SaimAMELX  ---------- ---------- ---------- -----..... .......... ..........
T65AMELX   ---------- ---------- ---------- -----...t. .......... ..........
A09_ANb    ---------- ---------- ---------- -----..... .......... ..........
A09_AMc    ---------- ---------- ---------- -----..... .......... ..........
SaimAMELY  ---------- ---------- ---------- -----..... .......... ...g......
M10_APb    ---------- ---------- ---------- -----..... .......... ..........
C71AMELY   ---------- ---------- ---------- -----..... .........a ..........
G72AMELY   ---------- ---------- ---------- -----..... .........a ..........
P11AMELY   ---------- ---------- ---------- -----..... .......... ..........
G73AMELY   ---------- ---------- ---------- -----..... .........a ..........
HsapAMELY  ---------- ---------- ---------- -----..... .........a ..........
PanAMELY   ---------- ---------- ---------- -----..... .........a ..........
LemurAMELX tgcagcctat ccagcccatc cagcccatcc agccc..... ....c..... ..........
OtoleAMELX tgcag----- ---------- ---cccatcc agccc..... .......... ..........
MusAMELX   ---------- ---------- ---------- -----...t. ....c....t ..........
EquusAMELX ---------- ---------- ---------- -----..... ....c..... .....c....
EquusAMELY ---------- ---------- ---------- -----..... ....a..... .....c....
AllExons   ---------- ---------- ---------- -----..... .......... ..........
```

```
[          7777777777 7777777777 7777777777 7777777777 7777777777 7777777777 ]
[          2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  ccctgccgcc acagccacct ctgcctccga tgtt-cccca tgcagcccct gcctcccatg
PanAMELX   .......... .......... .......... ....-..... .......... ..........
G73AMELX   .......... .......... .......... ....-..... .......... ..........
M10_AOa    .......... .......... .......... ....-..... .......... ..........
P70AMELX   .......... .......... .......... ....-..... .......... ..........
C71AMELX   ......t... .......... .......... ....-..... .......... ...c.....a
SaimAMELX  .......... .......... .......... ....-..... .......... ..........
T65AMELX   .......... .......... .......... ....-..... .......... ..........
A09_ANb    .......... .......... .......... ....-..... .......... ..........
A09_AMc    .......... .......... .......... ....-..... .......... ..........
SaimAMELY  .......... .......... .......... ....-..... .......... ..........
M10_APb    ......t... .......... .......... ....-..... .......... ...c.....a
C71AMELY   ......t... .......... ........a. ....-..... .......... ..........
G72AMELY   ......t... .......... .......... ....-..... .......... ...c.....a
P11AMELY   ......t... .......... ........a. ....-..... .......... ...c.g...a
G73AMELY   ......t... .......... ........a. ....-..... .a.g...... ...c.....a
HsapAMELY  ......t... .......... ........a. ....-....c ...g...... ...c.....a
PanAMELY   ......t... .......... ........a. ....-..... ...g...... ...c.....a
LemurAMELX .......... ........a. .........c ....-..... .......... ...g......
OtoleAMELX .......... .......... .........c ....-..... .......... ...g......
MusAMELX   .....g.a.. .......... ........ac ....-.t... .......... .t.c.....t
EquusAMELX .......... .......... .......... .a..-....t ........t. ...c..t...
EquusAMELY ..t....a.. ....g.t... ........a. .a..t....t .......t.. .g.c......
AllExons   .......... .......... .......... ....-..... .......... ..........

[          7777777777 7777777778 8888888888 8888888888 8888888888 8888888888 ]
[          8888888889 9999999990 0000000001 1111111112 2222222223 3333333334 ]
[          1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 ]
HsapAMELX  cttcctgatc tgactctgga agcttggcca tcaacagaca agaccaagcg ggaggaagtg
PanAMELX   .......... .......... .......... .......... .........- ----------
G73AMELX   .......... .......... .......... .......... .........a ..........
M10_AOa    .......... .......... .......... .....g.... .........a ..........
P70AMELX   .......... .......... .......... .....g.... .........a ..........
C71AMELX   .......... .......... .......... g......... .........a ..........
SaimAMELX  .......... ..c....... .......... g.....---- ---------- ----------
T65AMELX   .c........ ..c....... .......... g......... .........a ..........
A09_ANb    .......... ..c....... .......... g......... .........a ..........
A09_AMc    .......... ..c....... .......... g......... .........a ..........
SaimAMELY  .......... ..c-...... .......... g...------ ---------- ----------
M10_APb    .......... .......... .......... g......... .........a ..........
C71AMELY   .......... .......... .......... .....g.... .........a ...-------
G72AMELY   .......... ..ca...... .......... g......... .........a ..........
P11AMELY   .......... ..ca...... .......... g......... .......... ..........
G73AMELY   .......... ..ca...... .......... g......... .........a ..........
HsapAMELY  .......... ..ca...... .......... g......... .........a .........-
PanAMELY   .......... ..ca...... .......... g......... .......... ...a------
LemurAMELX .....c..c. ..c....... .......... g....t---- ---------- ----------
OtoleAMELX .....c.... ..c....... .......... g....g---- ---------- ----------
MusAMELX   ........g. ..c....... .......... g.g....... .......... ...a......
EquusAMELX ........c. ..c....... .......... g.......ac .a-------- ----------
EquusAMELY .....----- ---------- ---------- ---------- ---------- ----------
AllExons   .......... .......... .......... .......... .......... ..........
;
end;
```

VIII REFERENCES

Agulnik, A. I., A. Zharkikh, H. Boettger-Tong, T. Bourgeron, K. McElreavey, and C. E. Bishop. 1998. Evolution of the DAZ gene family suggests that Y-linked DAZ plays little, or a limited, role in spermatogenesis but underlines a recent African origin for human populations. Human Molecular Genetics **7**:1371-1377.

Aitken, R. J., and J. A. Marshall Graves. 2002. The future of sex. Nature **415**:963.

Alvesalo, L. 1997. Sex chromosomes and human growth. A dental approach. Human Genetics **101**:1-5.

Alvesalo, L., and P. Portin. 1980. 47,XXY males: sex chromosomes and tooth size. American Journal of Human Genetics **32**:955-959.

Andolfatto, P. 2001. Adaptive hitchhiking effects on genome variability. Current Opinion in Genetics and Development **11**:635-641.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution **18**:1585-1592.

Anisimova, M., R. Nielsen, and Z. Yang. 2003. Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. Genetics **164**:1229-1236.

Atanassov, I., C. Delichere, D. A. Filatov, D. Charlesworth, I. Negrutiu, and F. Moneger. 2001. Analysis and evolution of two functional Y-linked loci in a plant sex chromosome system. Molecular Biology and Evolution **18**:2162-2168.

Bachtrog, D. 2003. Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in Drosophila. Nature Genetics **34**:215-219.

Bachtrog, D. 2004. Evidence that positive selection drives Y-chromosome degeneration in Drosophila miranda. Nature Genetics **36**:518-522. Epub 2004 Apr 2025.

Bachtrog, D., and B. Charlesworth. 2002. Reduced adaptation of a non-recombining neo-Y chromosome. Nature **416**:323-326.

Baird, D. M., J. Coleman, Z. H. Rosser, and N. J. Royle. 2000. High levels of sequence polymorphism and linkage disequilibrium at the telomere of 12q: implications for telomere biology and human evolution. American Journal of Human Genetics **66**:235-250.

Baird, D. M., and N. J. Royle. 1997. Sequences from higher primates orthologous to the human Xp/Yp telomere junction region reveal gross rearrangements and high levels of divergence. Human Molecular Genetics **6**:2291-2299.

Bauer, V. L., and C. F. Aquadro. 1997. Rates of DNA sequence evolution are not sex-biased in Drosophila melanogaster and D. simulans. Molecular Biology and Evolution **14**:1252-1257.

Bennetzen, J. L., and B. D. Hall. 1982. Codon selection in yeast. Journal of Biological Chemistry **257**:3026-3031.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. 2002. GenBank. Nucleic Acids Research **30**:17-20.

Benton, M. J. 2000. Vertebrate Palaeontology. Blackwell Science, Oxford, Uk.

Berta, P., J. R. Hawkins, A. H. Sinclair, A. Taylor, B. L. Griffiths, P. N. Goodfellow, and M. Fellous. 1990. Genetic evidence equating SRY and the testis-determining factor. Nature **348**:448-450.

Betancourt, A. J., and D. C. Presgraves. 2002. Linkage limits the power of natural selection in Drosophila. Proceedings of the National Academy of Sciences of the United States of America **99**:13616-13620. Epub 12002 Oct 13607.

Bielawski, J. P., and Z. Yang. 2001. Positive and negative selection in the DAZ gene family. Molecular Biology and Evolution **18**:523-529.

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

Bohossian, H. B., H. Skaletsky, and D. C. Page. 2000. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. Nature **406**:622-625.

Brace, C. L. 1963. Structural Reduction in Evolution. American Naturalist **97**:39-49.

Bull, J. J. 1983. Evolution of sex determining mechanisms. The Benjamin/Cummings Publishing Company, Menlo Park, California.

Burgoyne, P. S. 1982. Genetic homology and crossing over in the X and Y chromosomes of Mammals. Human Genetics **61**:85-90.

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

Caballero, A. 1995. On the Effective Size of Populations With Separate Sexes, With Particular Reference to Sex-Linked Genes. Genetics **139**:1007-1011.

Carmichael, A. N., A. K. Fridolfsson, J. Halverson, and H. Ellegren. 2000. Male-biased mutation rates revealed from Z and W chromosome-linked ATP synthase alpha-subunit (ATP5A1) sequences in birds. Journal of Molecular Evolution **50**:443-447.

Carvalho, A. B. 2002. Origin and evolution of the Drosophila Y chromosome. Current Opinion in Genetics and Development **12**:664-668.

Carvalho, A. B., B. A. Dobo, M. D. Vibranovski, and A. G. Clark. 2001. Identification of five new genes on the Y chromosome of Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America **98**:13225-13230.

Carvalho, A. B., B. P. Lazzaro, and A. G. Clark. 2000. Y chromosomal fertility factors kl-2 and kl-3 of Drosophila melanogaster encode dynein heavy chain polypeptides. Proceedings of the National Academy of Sciences of the United States of America **97**:13239-13244.

Casane, D., S. Boissinot, B. H. J. Chang, L. C. Shimmin, and W. H. Li. 1997. Mutation pattern variation among regions of the primate genome. Journal of Molecular Evolution **45**:216-226.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov. 2002. Selection for short introns in highly expressed genes. Nature Genetics **31**:415-418. Epub 2002 Jul 2022.

Ceplitis, H., and H. Ellegren. 2004. Adaptive molecular evolution of HINTW, a female-specific gene in birds. Molecular Biology and Evolution **21**:249-254. Epub 2003 Aug 2029.

Chang, Y. M., L. A. Burgoyne, and K. Both. 2003. Higher failures of amelogenin sex test in an Indian population group. Journal of Forensic Science **48**:1309-1313.

Charlesworth, B. 1996. The evolution of chromosomal sex determination and dosage compensation. Current Biology **6**:149-162.

Charlesworth, B., and D. Charlesworth. 1978. A model for the evolution of dioecy and gynodioecy. American Naturalist **112**:975-997.

Charlesworth, B., and D. Charlesworth. 2000. The degeneration of Y chromosomes. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **355**:1563-1572.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics **134**:1289-1303.

Charlesworth, D., B. Charlesworth, and M. T. Morgan. 1995. The pattern of neutral molecular variation under the background selection model. Genetics **141**:1619-1632.

Charlesworth, D., M. T. Morgan, and B. Charlesworth. 1992. The effect of linkage and population size on inbreeding depression due to mutational load. Genetical Research **59**:49-61.

Chen, F. C., and W. H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. American Journal of Human Genetics **68**:444-456.

Chen, S. T. H., and W. S. B. Yeung. 1983. Sex control and sex reversal in fish under natural conditions. Fish Physiology **9**:171-222.

Ciccodicola, A., M. D'Esposito, T. Esposito, F. Gianfrancesco, C. Migliaccio, M. G. Miano, M. R. Matarazzo, M. Vacca, A. Franze, M. Cuccurese, M. Cocchia, A. Curci, A. Terracciano, A. Torino, S. Cocchia, G. Mercadante, E. Pannone, N. Archidiacono, M. Rocchi, D. Schlessinger, and M. D'Urso. 2000. Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. Human Molecular Genetics **9**:395-401.

Clark, A. G. 1997. Neutral behavior of shared polymorphism. Proceedings of the National Academy of Sciences of the United States of America **94**:7730-7734.

Comeron, J. M. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics **167**:1293-1304.

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

Crow, J. F. 2000a. A new study challenges the current belief of a high human male : female mutation ratio. Trends in Genetics **16**:525-526.

Crow, J. F. 2000b. The origins patterns and implications of human spontaneous mutation. Nature Reviews Genetics **1**:40-47.

Delbridge, M. L., and J. A. Graves. 1999. Mammalian Y chromosome evolution and the male-specific functions of Y chromosome-borne genes. Reviews of Reproduction **4**:101-109.

Delichere, C., J. Veuskens, M. Hernould, N. Barbacar, A. Mouras, I. Negrutiu, and F. Moneger. 1999. SlY1, the first active gene cloned from a plant Y chromosome, encodes a WD-repeat protein. Embo Journal **18**:4169-4179.

Desfeux, C., S. Maurice, J. P. Henry, B. Lejeune, and P. H. Gouyon. 1996. Evolution of reproductive systems in the genus Silene. Proceedings of the Royal Society of London Series B-Biological Sciences **263**:409-414.

Disteche, C. M. 1995. Escape from X inactivation in human and mouse. Trends Genet **11**:17-22.

Duan, J., M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Human Molecular Genetics **12**:205-216.

Ellegren, H. 2002. Dosage compensation: do birds do it as well? Trends in Genetics **18**:25-28.

Ellegren, H. 2000. Evolution of the avian sex chromosomes and their role in sex determination. Trends in Ecology & Evolution **15**:188-192.

Ellegren, H., and A. Carmichael. 2001. Multiple and independent cessation of recombination between avian sex chromosomes. Genetics **158**:325-331.

Ellegren, H., and A. K. Fridolfsson. 2003. Sex-specific mutation rates in salmonid fish. Journal of Molecular Evolution **56**:458-463.

Ellis, N. 1998. The war of the sex chromosomes. Nature Genetics **20**:9-10.

Ellis, N., P. Yen, K. Neiswanger, L. J. Shapiro, and P. N. Goodfellow. 1990. Evolution of the pseudoautosomal boundary in Old World monkeys and great apes. Cell **63**:977-986.

Eyre-Walker, A. 1993. Recombination and mammalian genome evolution. Proc R Soc Lond B Biol Sci **252**:237-243.

Eyre-Walker, A., and L. D. Hurst. 2001. The evolution of isochores. Nature Reviews Genetics **2**:549-555.

Felsenstein, J. 1974. The evolutionary advantage of recombination. Genetics **78**:737-756.

Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics **22**:521-565.

Filatov, D. A. 2002. ProSeq: A software for preparation and evolutionary analysis of DNA sequence data sets. Molecular Ecology Notes **2**:621-624.

Filatov, D. A. 2004. A gradient of silent substitution rate in the human pseudoautosomal region. Molecular Biology and Evolution **21**:410-417. Epub 2003 Dec 2005.

Filatov, D. A., and D. Charlesworth. 2002. Substitution Rates in the X- and Y-Linked Genes of the Plants, Silene latifolia and S. dioica. Molecular Biology and Evolution **19**:898-907.

Filatov, D. A., V. Laporte, C. Vitte, and D. Charlesworth. 2001. DNA diversity in sex-linked and autosomal genes of the plant species Silene latifolia and Silene dioica. Molecular Biology and Evolution **18**:1442-1454.

Filatov, D. A., F. Moneger, I. Negrutiu, and D. Charlesworth. 2000. Low variability in a Y-linked plant gene and it's implications for Y-chromosome evolution. Nature **404**:388-390.

Fredga, K. 1988. Aberrant chromosomal sex-determining mechanisms in mammals, with special reference to species with XY females. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **322**:83-95.

Fridolfsson, A. K., and H. Ellegren. 2000. Molecular evolution of the avian CHD1 genes on the Z and W sex chromosomes. Genetics **155**:1903-1912.

Garcia-Moreno, J., and D. P. Mindell. 2000. Rooting a Phylogeny with Homologous Genes on Opposite Sex Chromosomes (Gametologs): A Case Study Using Avian CHD. Molecular Biology and Evolution **17**:1826-1832.

Girondot, M., and J. Y. Sire. 1998. Evolution of the amelogenin gene in toothed and toothless vertebrates. European Journal of Oral Science **106**:501-508.

Glaser, B., D. Myrtek, Y. Rumpler, K. Schiebel, M. Hauwy, G. A. Rappold, and W. Schempp. 1999. Transposition of SRY into the ancestral pseudoautosomal region creates a new pseudoautosomal boundary in a progenitor of simian primates. Human Molecular Genetics **8**:2071-2078.

Glazko, G. V., and M. Nei. 2003. Estimation of divergence times for major lineages of primate species. Molecular Biology and Evolution **20**:424-434.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution **11**:725-736.

Gordo, I., and B. Charlesworth. 2001. The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes. Genetical Research **78**:149-161.

Gordo, I., and B. Charlesworth. 2000. On the speed of Muller's ratchet. Genetics **156**:2137-2140.

Gutman, D. S., and D. Charlesworth. 1998. An X-linked gene with a degenerate Y-linked homologue in a dioecious plant. Nature **393**:263-266.

Guttman, D. S., and D. Charlesworth. 1998. An X-linked gene with a degenerate Y-linked homologue in a dioecious plant. Nature **393**:263-266.

Haldane, J. B. S. 1947. The mutation rate of the gene for hemophilia and its segregation ratios in males and females. Annals of Eugenics **13**:262-271.

Hale, D. W. 1994. Is X-Y recombination necessary for spermatocyte survival during mammalian spermatogenesis? Cytogenetics and Cell Genetics **65**:278-282.

Hart, P. S., M. J. Aldred, P. J. Crawford, N. J. Wright, T. C. Hart, and J. T. Wright. 2002. Amelogenesis imperfecta phenotype-genotype correlations with two amelogenin gene mutations. Archives of Oral Biology **47**:261-265.

Hellborg, L., and H. Ellegren. 2004. Low levels of nucleotide diversity in mammalian Y chromosomes. Molecular Biology and Evolution **21**:158-163. Epub 2003 Oct 2031.

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Paabo, and M. Przeworski. 2003. A neutral explanation for the correlation of diversity with

recombination rates in humans. American Journal of Human Genetics **72**:1527-1535.
Hey, J., and R. M. Kliman. 2002. Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics **160**:595-608.
Hill, W. G., and A. Robertson. 1966. The effect of linkage on limits to artificial selection. Genetical Research **8**:269-294.
Hillis, D. M., and D. M. Green. 1990. Evolutionary changes of heterogametic sex in the phylogenetic history of amphibians. Journal of Evolutionary Biology **3**:49-64.
Hillson, S. 1996. Dental Anthropology. Cambridge University Press, Cambridge, UK.
Huang, W., B. H. Chang, X. Gu, D. Hewett-Emmett, and W. Li. 1997. Sex differences in mutation rate in higher primates estimated from AMG intron sequences. Journal of Molecular Evolution **44**:463-465.
Hudson, R. R., and N. L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111**:147-164.
Hurst, L. D., and H. Ellegren. 1998. Sex biases in the mutation rate. Trends in Genetics **14**:446-452.
Husmeier, D., and G. McGuire. 2003. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. Molecular Biology and Evolution **20**:315-337.
Husmeier, D., and F. Wright. 2001. Probabilistic divergence measures for detecting interspecies recombination. Bioinformatics **17**:S123-131.
Iida, K., and H. Akashi. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene **261**:93-105.
Iwase, M., Y. Satta, Y. Hirai, H. Hirai, H. Imai, and N. Takahata. 2003. The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. Proceedings of the National Academy of Sciences of the United States of America **100**:5258-5263.
Iwase, M., Y. Satta, and N. Takahata. 2001. Sex-chromosomal differentiation and amelogenin genes in mammals. Molecular Biology and Evolution **18**:1601-1603.
Jegalian, K., and D. C. Page. 1998. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. Nature **394**:776-780.
Kaessmann, H., V. Wiebe, G. Weiss, and S. Paabo. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nature Genetics **27**:155-156.
Kim, H. S., and O. Takenaka. 1996. A comparison of TSPY genes from Y-chromosomal DNA of the great apes and humans: sequence, evolution, and phylogeny. American Journal of Physical Anthropology **100**:301-309.
Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.
Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of Molecular Evolution **29**:170-179.

Kjellman, C., H.-O. Sjogren, and B. Widegren. 1995. The Y chromosome: a graveyard for endogenous retroviruses. Gene **161**:163-170.

Kliman, R. M., and J. Hey. 1993. Reduced natural selection associated with low recombination in Drosophila melanogaster. Molecular Biology and Evolution **10**:1239-1258.

Kondo, M., E. Nagao, H. Mitani, and A. Shima. 2001. Differences in recombination frequencies during female and male meioses of the sex chromosomes of the medaka, Oryzias latipes. Genetical Research **78**:23-30.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, and K. Stefansson. 2002. A high-resolution recombination map of the human genome. Nature Genetics **31**:241-247.

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

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

Kumar, S., and S. Subramanian. 2002. Mutation rates in mammalian genomes. Proceedings of the National Academy of Sciences of the United States of America **99**:803-808. Epub 2002 Jan 2015.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244-1245.

Kuroda, Y., N. Arai, M. Arita, M. Teranishi, T. Hori, M. Harata, and S. Mizuno. 2001. Absence of Z-chromosome inactivation for five genes in male chickens. Chromosome Research **9**:457-468.

Lagerstrom, M., N. Dahl, Y. Nakahori, Y. Nakagome, B. Backman, U. Landegren, and U. Pettersson. 1991. A deletion in the amelogenin gene (AMG) causes X-linked amelogenesis imperfecta (AIH1). Genomics **10**:971-975.

Lahn, B. T., and D. C. Page. 1999. Four evolutionary strata on the human X chromosome. Science **286**:964-967.

Lahn, B. T., and D. C. Page. 1997. Functional coherence of the human Y chromosome. Science **278**:675-680.

Lahn, B. T., N. M. Pearson, and K. Jegalian. 2001. The human Y chromosome, in the light of evolution. Nature Reviews Genetics **2**:207-216.

Lawson, L. J., and G. M. Hewitt. 2002. Comparison of substitution rates in ZFX and ZFY introns of sheep and goat related species supports the hypothesis of male- biased mutation rates. Journal of Molecular Evolution **54**:54-61.

Lebel-Hardenack, S., E. Hauser, T. F. Law, J. Schmid, and S. R. Grant. 2002. Mapping of Sex Determination Loci on the White Campion (Silene latifolia) Y Chromosome Using Amplified Fragment Length Polymorphism. Genetics **160**:717-725.

Lengerova, M., R. C. Moore, S. R. Grant, and B. Vyskot. 2003. The sex chromosomes of Silene latifolia revisited and revised. Genetics **165**:935-938.

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

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

Lercher, M. J., E. J. B. Williams, and L. D. Hurst. 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. Molecular Biology and Evolution **18**:2032-2039.

Li, W. H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. Journal of Molecular Evolution **24**:337-345.

Li, W. H., C. I. Wu, and C. C. Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular Biology and Evolution **2**:150-174.

Lien, S., J. Szyda, B. Schechinger, G. Rappold, and N. Arnheim. 2000. Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. American Journal of Human Genetics **66**:557-566.

Machado, C. A., R. M. Kliman, J. A. Markert, and J. Hey. 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of Drosophila pseudoobscura and close relatives. Molecular Biology and Evolution **19**:472-488.

Makalowski, W., and M. S. Boguski. 1998. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. Journal of Molecular Evolotion **47**:119-121.

Makova, K. D., and W. H. Li. 2002. Strong male-driven evolution of DNA sequences in humans and apes. Nature **416**:624-626.

Malcom, C. M., G. J. Wyckoff, and B. T. Lahn. 2003. Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. Molecular Biology and Evolution **20**:1633-1641. Epub 2003 Jul 1628.

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

Mathur, A. K., and P. D. Polly. 2000. The Evolution of Enamel Microstructure: How Important Is Amelogenin? Journal of Mammalian Evolution **7**:23-42.

Matsunaga, S., E. Isono, E. Kejnovsky, B. Vyskot, J. Dolezel, S. Kawano, and D. Charlesworth. 2003. Duplicative transfer of a MADS box gene to a plant Y chromosome. Molecular Biology and Evolution **20**:1062-1069. Epub 2003 Apr 1025.

May, C. A., A. C. Shone, L. Kalaydjieva, A. Sajantila, and A. J. Jeffreys. 2002. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. Nature Genetics **31**:272-275.

McGuire, G., and F. Wright. 2000. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. Bioinformatics **16**:130-134.

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

McVean, G. A., and B. Charlesworth. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155**:929-944.

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

Milne, I., F. Wright, G. Rowe, D. F. Marshall, D. Husmeier, and G. McGuire. 2004. TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments. Bioinformatics **20**:1806-1807. Epub 2004 Feb 1826.

Miyata, T., H. Hayashida, K. Kuma, K. Mitsuyasu, and T. Yasunaga. 1987. Male-driven molecular evolution: A model and nucleotide sequence analysis. Cold Spring Harbor Symposia on Quantitative Biology **52**:863-867.

Moore, R. C., O. Kozyreva, S. Lebel-Hardenack, J. Siroky, R. Hobza, B. Vyskot, and S. R. Grant. 2003. Genetic and functional analysis of DD44, a sex-linked gene from the dioecious plant Silene latifolia, provides clues to early events in sex chromosome evolution. Genetics **163**:321-334.

Mouchiroud, D., C. Gautier, and G. Bernardi. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. Journal of Molecular Evolution **40**:107-113.

Muller, H. J. 1932. Some genetic aspects of sex. American Naturalist **66**:118-138.

Nachman, M. W. 2001. Single nucleotide polymorphisms and recombination rate in humans. Trends in Genetics **17**:481-485.

Nachman, M. W. 1998. Y chromosome variation of mice and men. Molecular Biology and Evolution **15**:1744-1750.

Nachman, M. W., V. L. Bauer, S. L. Crowell, and C. F. Aquadro. 1998. DNA Variability and Recombination Rates at X-Linked Loci in Humans. Genetics **150**:1133-1141.

Nachman, M. W., and S. L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics **156**:297-304.

Nanda, I., U. Hornung, M. Kondo, M. Schmid, and M. Schartl. 2003. Common spontaneous sex-reversed XX males of the medaka Oryzias latipes. Genetics **163**:245-251.

Nanda, I., Z. Shan, M. Schartl, D. W. Burt, M. Koehler, H.-G. Nothwang, F. Grutzner, I. R. Paton, D. Windsor, I. Dunn, W. Engel, P. Staeheli, S. Mizuno, T. Haaf, and M. Schmid. 1999. 300 million years of conserved synteny between chicken Z and human chromosome 9. Nature Genetics **21**:258-259.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the number of synonymous and non-synonymous nucleotide substitutions. Molecular Biology and Evolution **3**:418-426.

Obara, M., S. Matsunaga, S. Nakao, and S. Kawano. 2002. A plant Y chromosome-STS marker encoding a degenerate retrotransposon. Genes and Genetic Systems **77**:393-398.

Ogawa, A., K. Murata, and S. Mizuno. 1998. The location of Z- and W-linked marker genes and sequence on the homomorphic sex chromosomes of the ostrich and the emu. Proceedings of the National Academy of Sciences of the United States of America **95**:4415-4418.

Ogawa, Y., and J. T. Lee. 2002. Antisense regulation in X inactivation and autosomal imprinting. Cytogenetic and Genome Research **99**:59-65.

Orr, H. A., and Y. Kim. 1998. An adaptive hypothesis for the evolution of the Y chromosome. Genetics **150**:1693-1698.

Pal, C., and L. D. Hurst. 2003. Evidence for co-evolution of gene order and recombination rate. Nature Genetics **33**:392-395.

Pamilo, P., and N. O. Bianchi. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol **10**:271-281.

Peck, J. R. 1994. A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics **137**:597-606.

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

Pecon Slattery, J., L. Sanner-Wachter, and S. J. O'Brien. 2000. Novel gene conversion between X-Y homologues located in the nonrecombining region of the Y chromosome in Felidae (Mammalia). Proceedings of the National Academy of Sciences of the United States of America **97**:5307-5312.

Peichel, C. L., J. A. Ross, C. K. Matson, M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, S. Mori, D. Schluter, and D. M. Kingsley. 2004. The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. Current Biology **14**:1416-1424.

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

Pigozzi, M. I., and A. J. Solari. 1997. Extreme axial equalization and wide distribution of recombination nodules in the primitive ZW pair of Rhea americana (Aves, Ratitae). Chromosome Research **5**:421-428.

Plotkin, J. B., H. Robins, and A. J. Levine. 2004. Tissue-specific codon usage and the expression of human genes. Proceedings of the National Academy of Sciences of the United States of America **101**:12588-12591. Epub 12004 Aug 12516.

Posada, D., and K. A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proceedings of the National Academy of Sciences of the United States of America **98**:13757-13762.

Purvis, A. 1995. A composite estimate of primate phylogeny. Philosophical Transactions of the Royal Society - B: Biological Sciences **348**:405-421.

Rao, E., B. Weiss, M. Fukami, A. Rump, B. Niesler, A. Mertz, K. Muroya, G. Binder, S. Kirsch, M. Winkelmann, G. Nordsiek, U. Heinrich, M. H. Breuning, M. B. Ranke, A. Rosenthal, T. Ogata, and G. A. Rappold.

1997. Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. Nature Genetics **16**:54-63.

Rappold, G. A. 1993. The pseudoautosomal regions of the human sex chromosomes. Human Genetics **92**:315-324.

Reeve, H. K., and D. W. Pfennig. 2003. Genetic biases for showy males: are some genetic systems especially conducive to sexual selection? Proceedings of the National Academy of Sciences of the United States of America **100**:1089-1094.

Reich, D. E., S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. Nature Genetics **32**:135-142.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics **16**:276-277.

Rice, W. R. 1996. Evolution of the Y sex chromosome in animals. Bioscience **46**:331-343.

Rice, W. R. 1987. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. Genetics **116**:161-167.

Rice, W. R. 1994. Degeneration of a nonrecombining chromosome. Science **263**:230-232.

Rice, W. R. 1992. Sexually antagonistic genes: experimental evidence. Science **256**:1436-1439.

Rodriguez, I. R., K. Mazuruk, T. J. Schoen, and G. J. Chader. 1994. Structural analysis of the human hydroxyindole-O-methyltransferase gene. Presence of two distinct promoters. Journal of Biological Chemistry **269**:31969-31977.

Rouyer, F., M. C. Simmler, C. Johnsson, G. Vergnaud, H. J. Cooke, and J. Weissenbach. 1986. A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes. Nature **319**:291-295.

Rozas, J., and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15**:174-175.

Salido, E. C., P. H. Yen, K. Koprivnikar, L. C. Yu, and L. J. Shapiro. 1992. The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes. American Journal of Human Genetics **50**:303-316.

Sandstedt, S. A., and P. K. Tucker. 2004. Evolutionary strata on the mouse X chromosome correspond to strata on the human X chromosome. Genome Research **14**:267-272.

Schiebel, K., J. Meder, A. Rump, A. Rosenthal, M. Winkelmann, C. Fischer, T. Bonk, A. Humeny, and G. Rappold. 2000. Elevated DNA sequence diversity in the genomic region of the phosphatase PPP2R3L gene in the human pseudoautosomal region. Cytogenetics and Cell Genetics **91**:224-230.

Seffens, W., and D. Digby. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Research **27**:1578-1584.

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden. 1995. DNA sequence evolution: the sounds of silence. Philosophical

Transactions of the Royal Society of London Series B-Biological Sciences **349**:241-247.

Sharp, P. M., and W. H. Li. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution **24**:28-38.

Shen, L. X., J. P. Basilion, and V. P. Stanton, Jr. 1999. Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proceedings of the National Academy of Sciences of the United States of America **96**:7871-7876.

Shen, P., F. Wang, P. A. Underhill, C. Franco, W. H. Yang, A. Roxas, R. Sung, A. A. Lin, R. W. Hyman, D. Vollrath, R. W. Davis, L. L. Cavalli-Sforza, and P. J. Oefner. 2000. Population genetic implications from sequence variation in four Y chromosome genes. Proceedings of the National Academy of Sciences of the United States of America **97**:7354-7359.

Shields, D. C., P. M. Sharp, D. G. Higgins, and F. Wright. 1988. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Molecular Biology and Evolution **5**:704-716.

Shimmin, L. C., B. H. Chang, and W. H. Li. 1993. Male-driven evolution of DNA sequences. Nature **362**:745-747.

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

Singer, A., H. Perlman, Y. Yan, C. Walker, G. Corley-Smith, B. Brandhorst, and J. Postlethwait. 2002. Sex-specific recombination rates in zebrafish (Danio rerio). Genetics **160**:649-657.

Siroky, J., M. R. Castiglione, and B. Vyskot. 1998. DNA methylation patterns of Melandrium album chromosomes. Chromosome Research **6**:441-446.

Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S. F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S. P. Yang, R. H. Waterston, R. K. Wilson, S. Rozen, and D. C. Page. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423**:825-837.

Smith, J. M., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genetical Research **23**:23-35.

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

Smith, N. G. C., and L. D. Hurst. 1999. The causes of synonymous rate variation in the rodent genome: Can substitution rate be used to estimate the sex bias in mutation rate? Genetics **152**:661-673.

Staden, R., K. F. Beal, and J. K. Bonfield. 2000. The Staden package, 1998. Methods in Molecular Biology **132**:115-130.

Strathern, J. N., B. K. Shafer, and C. B. McGill. 1995. DNA-Synthesis Errors Associated with Double-Strand-Break Repair. Genetics **140**:965-972.

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

Swanson, W. J., R. Nielsen, and Q. Yang. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Molecular Biology and Evolution **20**:18-20.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nature Reviews Genetics **3**:137-144.

Swofford, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinnauer Associates, Sunderland, Massachusetts.

Sykes, B. 2003. Adam's Curse - A Future Without Men. Bantam Press, London.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution **10**:512-526.

Taylor, D. R. 1994. Sex ratio in hybrids between Silene alba and Silene dioica: evidence for Y-linked restorers. Heredity **73**:518-526.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position- Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Research **22**:4673-4680.

Urrutia, A. O., and L. D. Hurst. 2003. The signature of selection mediated by expression on human genes. Genome Research **13**:2260-2264. Epub 2003 Sep 2215.

Van Doorn, G. S., P. C. Luttikhuizen, and F. J. Weissing. 2001. Sexual selection at the protein level drives the extraordinary divergence of sex-related genes during sympatric speciation. Proceedings of the Royal Society of London Series B-Biological Sciences **268**:2155-2161.

Wallace, H., G. M. Badawy, and B. M. Wallace. 1999. Amphibian sex determination and sex reversal. Cellular and Molecular Life Sciences **55**:901-909.

Weinberg, W. 1912. Zur Vererbung des zwergwuchses. Arch Rassen-u Gesel Biolog **9**:710-718.

Wolfe, K. H., and P. M. Sharp. 1993. Mammalian Gene Evolution - Nucleotide-Sequence Divergence between Mouse and Rat. Journal of Molecular Evolution **37**:441-456.

Wu, C. I., and W. H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proceedings of the National Academy of Sciences of the United States of America **82**:1741-1745.

Wyckoff, G. J., J. Li, and C. I. Wu. 2002. Molecular evolution of functional genes on the mammalian Y chromosome. Molecular Biology and Evolution **19**:1633-1636.

Wyckoff, G. J., W. Wang, and C.-I. Wu. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature **403**:304-309.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences **13**:555-556.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution **15**:568-573.

Yang, Z., and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution **19**:908-917.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

Yi, S., D. L. Ellsworth, and W. H. Li. 2002. Slow molecular clocks in Old World monkeys, apes, and humans. Molecular Biology and Evolution **19**:2191-2198.

Yi, S., T. J. Summers, N. M. Pearson, and W. H. Li. 2004. Recombination has little effect on the rate of sequence divergence in pseudoautosomal boundary 1 among humans and great apes. Genome Research **14**:37-43. Epub 2003 Dec 2012.

Yoder, A. D., and Z. Yang. 2000. Estimation of Primate Speciation Dates Using Local Molecular Clocks. Molecular Biology and Evolution **17**:1081-1090.

Yu, A., C. F. Zhao, Y. Fan, W. H. Jang, A. J. Mungall, P. Deloukas, A. Olsen, N. A. Doggett, N. Ghebranious, K. W. Broman, and J. L. Weber. 2001. Comparison of human genetic and sequence-based physical maps. Nature **409**:951-953.

Zhang, J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. Molecular Biology and Evolution **21**:1332-1339. Epub 2004 Mar 1310.

Zinn, A. R., D. C. Page, and E. M. Fisher. 1993. Turner syndrome: the case of the missing sex chromosome. Trends in Genetics **9**:90-93.