Molecular evolution under low recombination

Vera B. Kaiser

Submitted for the degree of PhD

The University of Edinburgh

2009

Declaration

This thesis was composed by myself, and is a presentation of my original research work, except where otherwise stated in the text.

The work was done under the guidance of Professor Brian Charlesworth, Dr. Roberta Bergero and Prof. Deborah Charlesworth at the University of Edinburgh. It has not been submitted for any other degree or professional qualification.

Vera Kaiser, November 22, 2009

Acknowledgements

I am very grateful to my supervisors Brian Charlesworth, Roberta Bergero and Deborah Charlesworth for their great effort and support throughout my PhD. This work certainly would not have been the same without them.

I would like to thank Nick Barton and John Brookfield for making my PhD viva a pleasurable experience, and for their useful feedback and advice on the thesis.

My thanks go to everyone in the Charlesworth Lab, for discussions, enjoyable journal clubs and a pleasurable working atmosphere in the department. Special thanks go to Andrea Betancourt, Kelly Dyer, John Welch and Laurence Loewe for explaining things to me – and to Beatriz Vicoso for sharing some of the burden of being a PhD student.

I am grateful to the University of Edinburgh for funding my PhD. The university's funding towards the many sports clubs and societies has certainly made this PhD a more enjoyable experience.

Last but not least, I would like to thank my family and friends for their support throughout.

Abstract

Analyzing regions in the genome with low levels of recombination helps understand the prevalence of sexual reproduction. Here, I show that variability in regions of reduced recombination in Drosophila can be explained by interference among strongly deleterious mutations; selection becomes progressively less effective in influencing the behaviour of neighbouring sites as the number of closely linked sites on a chromosome increases. I also show that the accumulation of loss-of-function mutations on the neo-Y chromosome of *Drosophila miranda* is compatible with a model of selection against such mutations alone, without the need to invoke the action of selective sweeps. I describe the discovery of two new sex-linked genes in the plant *Silene latifolia*, *SlCyt* and *SlX9/SlY9*. *SlCyt* has been recently translocated from an autosome to the X and shows signs of a selective sweep. Its possible role in having caused recombination arrest between the evolving X and Y chromosome is discussed. *SlX9* still has an intact Y-linked copy that is presumably functional. Nucleotide diversity at *SlY9* is very low, whereas *SlX9* has an unusually high diversity and shows signs of introgression from *S. dioica* into *S. latifolia*, but the effect of this seems very localized.

Contents

1 Introduction

1.1.1 Evolutionary consequences of reduced recombination

Recombination, the exchange of genetic material between homologous sites or chromosomes, occurs in most prokaryotic and eukaryotic organisms. However, it is still not fully understood why recombination is so prevalent, given that there are major costs associated with it (BARTON and CHARLESWORTH 1998; WEST *et al.* 1999). Most importantly, recombination leads to the loss of beneficial allele combinations in each generation, and, in organisms where recombination is linked to sexual reproduction, it is associated with a "two-fold cost of sex" since females also need to produce male offspring (MAYNARD SMITH 1978). Mechanistic explanations for the existence of recombination, such as the proper disjunction of homologous chromosomes during meiosis, cannot explain its universal prominence since some organisms, such as male Drosophila, can undergo meiosis without recombination (GETHMANN 1988), and some organisms do indeed reproduce without sex - even though these lineages tend to be short-lived on an evolutionary time scale (MAYNARD SMITH 1978). Instead, research has focused on describing the population genetics consequences of the absence of sex and recombination, and relating these models to observations from systems in which sex or recombination is indeed absent.

A fundamental fact is that, without recombination, evolution at any one site in the genome is not independent of evolution at other, linked sites. In particular, assuming that selection is acting to either remove deleterious mutations from the population, or to increase the frequency of beneficial alleles, selective processes acting on different variants simultaneously will not be independent. These processes of selective interference are collectively called Hill-Robertson Interference (HILL and ROBERTSON 1966). They occur because a single chromosome will necessarily not contain all advantageous variants that are segregating in the population at any one time and, without recombination, associations between good and bad alleles cannot be broken up. Without recombination, two advantageous mutations that occurred in different individuals cannot be united onto the same chromosome, nor can deleterious mutations be brought together and eliminated from the population simultaneously (FISHER 1930; MORGAN 1913; MULLER 1932); this increases the time of their segregation within the population and leads to the build-up of negative linkage disequilibrium among selected sites (an overrepresentation of good-bad associations) (FELSENSTEIN 1974; HILL and ROBERTSON 1966). Depending on the type and strength of selection, we need to deal with different models of the general Hill-Robertson effect.

1.1.2 Models of interference among selected sites

Selective sweeps occur when a beneficial mutation becomes fixed in a population due to strong positive selection (KAPLAN *et al.* 1989; MAYNARD SMITH and HAIGH 1974). If there is no recombination, a sweep will drag to fixation all linked variants that were present on the same chromosome. Hence, deleterious variants can become fixed by a sweep, provided that the selected mutation's positive impact on fitness outweighs the cost of all deleterious mutations linked to it (HADANY and FELDMAN 2005; JOHNSON and BARTON 2002). On the other hand, any beneficial mutation that was also present in the population before the sweep will be lost.

Another model describes the removal of strongly deleterious mutations from the population: background selection (CHARLESWORTH 1994; CHARLESWORTH *et al.* 1993; NORDBORG *et al.* 1996). In this model, strongly deleterious mutations entering the population are constantly removed by selection, and, when there is no recombination, all neutral or weakly selected sites which are linked to the deleterious mutation will be eliminated, too. In this respect, background selection is very much like a selective sweep, only that variants are being removed by selection instead of becoming fixed.

A third model, Muller's ratchet, describes the successive loss of the least-loaded mutational class from an asexual population (FELSENSTEIN 1974; MULLER 1964). Since mutations occur randomly, individuals vary in the number of deleterious mutations that they carry. If the population size is small, the class of individuals that carries the fewest number of deleterious mutations may become lost by chance. Without recombination, this class cannot be restored, i.e. the ratchet has made one "click", and there will be a new best class of individuals that can become lost again. Since, in this model, all mutations are assumed to be deleterious and irreversible, the fitness of the population decays continuously.

Finally, the "weak selection Hill-Robertson effect" describes interference among weakly selected alleles, which spend a long time segregating within the population before they become fixed or lost (COMERON *et al.* 2008; HILL and ROBERTSON 1966; MCVEAN and CHARLESWORTH 2000). Since two beneficial mutations cannot be combined onto the same chromosome, they will segregate simultaneously and impede each other's fixation.

1.1.3 The concept of effective population size

In the context of Hill-Robertson interference, the concept of the effective population size, N_e, plays a fundamental role: The effective population size is defined as the size of an idealized Wright-Fisher population that experiences the same amount of genetic drift, i.e. the same amount of random sampling of alleles, as the real population (CHARLESWORTH 2009; FISHER 1930; WRIGHT 1931). In a Wright-Fisher population, $N = N_e$ diploid hermaphrodites reproduce by random mating; gametes are produced "with replacement", i.e. each individual produces a Poisson number of offspring, there is free recombination, and discrete generations. Due to its finite size, the population experiences random fluctuations in allele frequencies (i.e. drift), and the rate of change in allele frequency is a function of $1/(2N)$. The effective population size is a very important concept because it determines levels of variability within a population, as well as the efficacy of selection: Neutral diversity, the average pairwise difference between nucleotide sites, is directly proportional to N_e ($\pi = 4 N_e\mu$) (KIMURA 1983), where μ is the neutral mutation rate per nucleotide site. The chance of fixation of a beneficial or deleterious mutation is also a function of N_e. In particular, the product of N_e times the selection coefficient ($N_e\, s$) determines the efficacy of selection as opposed to drift. When $N_e\, s > 1$, the fate of a mutation is largely determined by selection, whereas for $N_e\, s << 1$, it is mainly determined by drift, i.e. it is effectively neutral (CROW and KIMURA 1970; EWENS 2004). Hence, measuring neutral diversity in a natural population also gives information about the efficacy of selection because low levels of diversity indicate a lower N_e.

The effective population size varies between species, but also within a genome, since it is affected by selection at neighbouring sites. All sub-categories of Hill-Robertson interference have in common the property that they reduce the N_e of the surrounding genomic region (FELSENSTEIN 1974; HILL and ROBERTSON 1966). With reduced recombination, variants associated with a selected allele have an increased or decreased chance of contributing to the next generation compared to random sampling, increasing the variance in their reproductive success. Accordingly, fewer gametes will, on average, contribute to the next generation. Since the efficacy of selection depends on N_es, a reduction in N_e has the consequence that the chance for a beneficial mutation to go to fixation is reduced, whereas deleterious mutations have an increased chance of fixation (BIRKY and WALSH 1988; KIMURA 1983; ORR and KIM 1998). Hence, we expect lower levels of adaptation when recombination is reduced or absent, and regions of the genome that do not recombine are expected to degenerate over time.

Not only the type, but also the strength of selection differs in the different models of Hill-Robertson interference: selective sweeps and background selection occur when selection at the sites in question is strong ($N_es > 1$), so that there is a negligible chance for deleterious mutations to go to fixation (CHARLESWORTH 1994; CHARLESWORTH *et al.* 1993), whereas beneficial mutations become rapidly fixed during a sweep. Muller's ratchet works when selection against deleterious mutations is sufficiently strong to keep mutations in mutation-selection equilibrium, but weak enough to allow the occasional loss of the least-loaded class by drift (HAIGH 1978). "Weak selection" Hill-Robertson interference occurs when (positive or negative) selection is of the order of $N_es \sim 0$ to 1.0, so that alleles spend a long time segregating in the population before they become fixed or lost (MCVEAN and CHARLESWORTH 2000). Also the rate of occurrence of new mutations plays a role in the different models: for example, weak-selection interference will only occur if mutations are segregating simultaneously, setting a lower limit for their rate of mutation. Similarly, the rate at which Muller's ratchet leads to mutational meltdown crucially depends on the rate of mutation.

1.1.4 Evidence of a reduction in N_e due to interference effects

In the real world, rates of recombination vary and do so at different scales: the whole genome is inherited as a single linkage group in asexual species and in organelles (LYNCH and BLANCHARD 1998; MORAN 1996; SPRATT and MAIDEN 1999); in *Drosophila*, recombination is restricted to females and the fourth chromosome does not cross over at all (even though gene conversion might take place) (JENSEN *et al.* 2002; WANG *et al.* 2002); the evolution of sex chromosomes is characterized by the lack of recombination between the X and Y (CHARLESWORTH and CHARLESWORTH 2000; CHARLESWORTH 2002; CHARLESWORTH *et al.* 2005); recombination rates also vary along single chromosomes, generally being reduced near the centromere, and the study of mutational hotspots in humans showed that recombination rates evolve quickly and even show high within-species variation (MYERS *et al.* 2005). In line with theory, there is evidence for a reduction in diversity in regions of reduced recombination in *Drosophila* (ANDOLFATTO 2001; BEGUN and AQUADRO 1992; PRESGRAVES 2005), though the evidence is less clear-cut in other organisms (HELLMANN *et al.* 2005; WRIGHT *et al.* 2006). There is also evidence for reduced levels of adaptation when recombination is infrequent; for example, in Drosophila, there are reduced levels of codon usage bias and a higher rate of fixation of deleterious variants in regions of low recombination (BETANCOURT *et al.* 2009; HADDRILL *et al.* 2007). In Bdelloid rotifers, an ancient asexual clade, deleterious alleles segregate at higher frequencies compared to facultative sexual Daphnia or monogonont rotifers (BARRACLOUGH *et al.* 2007); similarly, the rate of amino acid to silent substitution (K_A/K_S) is increased in obligate asexual lineages compared to sexual lineages of Daphnia (PALAND and LYNCH 2006). The degeneration of non-recombining Y chromosomes, which will be discussed in more detail below, is a prime example of degeneration due to lack of recombination. Furthermore, theoretical studies have shown that a short-term evolutionary advantage of a modifier of recombination can exist under a variety of different scenarios where the lack of recombination reduces population mean fitness (BARTON and OTTO 2005; KEIGHTLEY and OTTO 2006; OTTO and BARTON 1997; ROZE and BARTON 2006).

1.1.5 Are deleterious mutations the main cause of degeneration?

Given the evidence for a reduction in diversity and adaptation in regions of low recombination, this still leaves the question of which processes have led to the situation. A reasonable assumption is that most organisms are well-adapted to their environment, and hence most new mutations at functional sites are likely to be deleterious. However, our knowledge of the rates of occurrence of beneficial and deleterious mutations, and their associated fitness effects, is still rather limited, making quantitative predictions difficult. Furthermore, positive or negative selection can leave similar traces in the genome by distorting the frequency spectrum of segregating sites in similar ways, making it hard to distinguish between the two causes (BACHTROG 2004; CHARLESWORTH and CHARLESWORTH 2000; CHARLESWORTH *et al.* 1995; KAPLAN *et al.* 1989).

Efforts have been made to better understand spontaneous rates of selected mutations as well as their effects on fitness. Mutation rates can be measured either directly in mutation-accumulation experiments, or by between-species comparisons of divergence at putatively neutral sites. Using divergence data, the amount of nonsynonymous changes that did not become fixed between species, relative to the amount of change at putative neutral sites, can be considered as the deleterious component. Typically, the vast majority of nonsynoymous changes are deleterious; there is on average more than one new deleterious mutation per individual in each generation in Drosophila and humans, especially if non-coding sequences are taken into account (EYRE-WALKER and KEIGHTLEY 1999; HAAG-LIAUTARD *et al.* 2007).

Deleterious mutations must vary in the effects that they have on fitness. One method to investigate the distribution of mutational effects is to compare patterns of polymorphisms and divergence between two closely related species that differ in their effective population sizes. Depending on the shape of the distribution of mutational effects, a different proportion of mutations will be effectively neutral or deleterious in the two species. Neutral diversity will be directly proportional to N_e, whereas variants under sufficiently strong selection will not differ much in abundance. This approach has led to the conclusion that, at least in Drosophila, the vast majority (about 90%) of nonsynonymous mutations are strongly deleterious, whereas a few percent are effectively neutral (LOEWE *et al.* 2006). Furthermore, the distribution of fitness effects that best incorporates silent diversity estimates and the frequency of lethal mutations in Drosophila is the log-normal distribution (LOEWE and CHARLESWORTH 2006), though this might not hold for other species (EYRE-WALKER and KEIGHTLEY 2007). For example, a gamma distribution of fitness effects is a good fit to the frequency spectrum of segregating nonsynonymous mutations in humans (EYRE-WALKER *et al.* 2006).

1.2.1 Using sex chromosomes to study the effects of the absence of recombination

Sex chromosomes are the ultimate test grounds for studying degenerative processes acting in regions of reduced recombination. In particular, in systems in which the X and the Y chromosome evolved from a pair of autosomes, it is possible to make a direct comparison between homologous genes situated in non-recombining regions on the Y versus those that still recombine on the X. In this section, I will refer to male heterogamety (males being XY and females XX), even though the same processes occur in female heterogamety (ZW/ZZ systems).

There seems to be a common route for most incipient sex chromosomes, irrespective of whether males or females are the heterogametic sex. Starting with a species that is entirely hermaphroditic or in which sex is determined by environmental cues, the evolution of separate sexes under genetic sex determination requires the presence of two primary sex-determining loci: one dominant female-suppressor, and one recessive male sterility locus (CHARLESWORTH and CHARLESWORTH 1978), both of which need to be polymorphic for sexually antagonistic alleles, i.e. alleles that are beneficial for one sex function and/or deleterious for the other (RICE 1984). Recombination between these sex-determining genes is then restricted to prevent the production of maladapted or infertile offspring (CHARLESWORTH and CHARLESWORTH 1978; NEI 1969), creating the first region on the sex chromosomes that only recombines in the homogametic sex (i.e. XX females). This region may expand due to the accumulation of sexually antagonistic, male-benefiting genes on the Y and/or chromosomal rearrangements (CHARLESWORTH and CHARLESWORTH 2000; CHARLESWORTH *et al.* 2005; RICE 1987). Eventually, recombination between the X and Y becomes restricted to the pseudoautosomal region or is completely absent. The non-recombining part of the Y chromosome tends to undergo a process of degeneration, typically retaining only few genes specific to male function (CHARLESWORTH and CHARLESWORTH 2000). The accumulation of transposable elements and non-coding sequence, such as and heterochromatin, is a further common feature of Y chromosome evolution (ERLANDSSON *et al.* 2000; STEINEMANN and STEINEMANN 2005), possibly to reduce the expression of non-functional genes. Dosage compensation may evolve to compensate for gene loss from the Y, equalizing expression levels between the X and the autosomes (CHARLESWORTH 1996).

The degeneration of evolving Y chromosomes can occur due to any types of interference among selected sites, though the relative contributions of positive versus negative selection are still under debate. Apart from models involving selection, the effective population size of the Y is also reduced, simply because there are only 1/4 as many Y chromosomes in the populations compared to the autosomes, and 1/3 as many compared to the X. Furthermore, if a deleterious mutation on the Y is recessive, it will be masked in males as it never occurs in the homozygous state, which increases its chance of fixation.

1.2.2 Examples of sex chromosomes

Sex chromosomes can be found throughout the animal and plant kingdom, where they emerged independently numerous times, e.g. in plants (CHARLESWORTH 2002), birds (LAWSON-HANDLEY *et al.* 2004), fish (PEICHEL *et al.* 2004), insects (SANCHEZ 2008) or mammals (SKALETSKY *et al.* 2003; VEYRUNES *et al.* 2008; WATERS *et al.* 2001), though not in fungi (which have mating-type loci with similar properties to sex chromosomes (FRASER *et al.* 2004), reviewed in BERGERO and CHARLESWORTH (2009)). The evolution of sex chromosomes is remarkably similar in different organisms and a reduction in recombination between the sex determining loci and subsequent Y chromosome degeneration usually plays a role. The age of a sex chromosome system can be estimated by comparing X- and Y-linked homologues, calculating their sequence divergence, and using a molecular clock to calibrate their time of divergence; alternatively, a phylogenetic comparison with related species that have different breeding systems can provide information on the age of the sex chromosomes.

The sex chromosomes of eutherian mammals evolved from an ordinary pair of autosomes about 170 MYA, and Y degeneration is very advanced (LAHN and PAGE 1999); whereas the X carries about 2000 genes, only 78 functional genes are left on the heterochromatic Y (SKALETSKY *et al.* 2003), some of which were translocated onto the Y from the autosomes, possibly because they are specifically advantageous to male function (RICE 1984). On the other extreme, sex chromosomes in papaya still recombine along most of their length, and recombination between the X and Y is only restricted in a small region where the sex determination loci are located (LIU *et al.* 2004). However, also in papaya, there are early signs of Y degeneration, such as the accumulation of repetitive sequences in Y-linked introns. There is recent evidence of an even younger pair of sex chromosomes in the wild strawberry, where recombination between the sex determining genes is reduced but not yet fully suppressed, so that, apart from hermaphrodites, neuter individuals are also produced; this must be very costly, and we anticipate strong selection to reduce recombination in the region. There are also very ancient sex chromosome systems, such as the Emu, where the W and Z still recombine along most of their length (OGAWA *et al.* 1998; SHETTY *et al.* 1999), even though the sex chromosomes are very old, having presumably arisen before the radiation of birds, about 120 MYA (VAN TUINEN and HEDGES 2001), and different stages of W degeneration can be found within snakes (MATSUBARA *et al.* 2006).

In *Drosophila melanogaster*, the Y chromosome is very small and heterochromatic, containing only about 20 genes (CARVALHO *et al.* 2009). Since there is no recombination between homologues in Drosophila males, there is no pseudoautosomal region on Drosophila sex chromosomes. The origin of the *D. melanogaster* Y is not fully known because there is no apparent homology with the X, and all genes identified on the Y seem to have arisen by duplication from the autosomes (BROSSEAU 1960; CARVALHO 2002). Hence, the original Y might have been completely lost, and replaced by a new Y chromosome, derived from an unrelated genetic element that acquired male function genes as well as the ability to segregate with the X (CARVALHO *et al.* 2009). Within the genus of Drosophila, several other large-scale changes of sex chromosomes have been identified, such as autosomal additions to existing X and Y chromosomes, creating neo-sex chromosome systems (CHARLESWORTH and CHARLESWORTH 2005). In one such case (*D. miranda*), an autosome arm became fused to the existing Y chromosome about 1.75 MYA, creating a neo-Y chromosome that stopped recombining instantaneously (because of the lack of recombination in Drosophila males) (BARTOLOME and CHARLESWORTH 2006; STEINEMANN and STEINEMANN 1998). The corresponding X chromosome did not fuse with the neo-X, but segregates from the neo-Y. Genes on the neo-Y and neo-X are clear homologues, and the neo-Y chromosome is currently undergoing a process of degeneration, having lost about half of its genes within a very short evolutionary time-scale (BACHTROG *et al.* 2008). Explaining the very high rate of accumulation of loss-of function mutations on the neo-Y will be the topic of chapter 3.

Additions of autosomal sequences onto sex chromosomes have also occurred in the history of the human sex chromosomes, possibly contributing to the evolution of evolutionary strata, i.e. regions with differential X-Y divergence. For example, one region of the human sex chromosomes, which shows very low X-Y divergence, is still autosomal in marsupials, suggesting that this region was added less than ~ 166 MYA (VEYRUNES *et al.* 2008; WATSON *et al.* 1991). Evolutionary strata might also be caused by chromosomal inversions or other mechanisms that lead to a cessation of recombination between the X and the Y, as discussed in chapter 4.

1.2.3 The model sex chromosome system of *Silene latifolia*

Most well-studied systems are old (Drosophila, mammals, birds), with degenerated and gene-poor Y (or W) chromosomes. When studying these older systems, it is hard to infer the evolutionary processes that occurred in the initial stages of X-Y differentiation, i.e. the mechanisms of how recombination suppression might have evolved, and the consequences of loss of recombination.

In chapters four and five, I use empirical methods to study the sex chromosomes of the plant species *Silene latifolia* (Caryophyllaceae), the white campion. *S. latifolia* has comparatively young sex chromosomes; dioecy, the state of having separate sexes, evolved only about 5-10 MYA (BERGERO *et al.* 2007; FILATOV 2005). Most of *S. latifolia's* close relatives are hermaphrodites that do not have sex chromosomes, and divergence between *S. latifolia* and the hermaphroditic *S. conica* and gynodioecious *S. vulgaris* (which contains both female and hermaphroditic individuals) is low (silent divergence values of about 15% or 20% respectively (FILATOV 2008; FILATOV and CHARLESWORTH 2002), setting an upper limit for the age of the system. In line with this, the maximal divergence between X-and Y-linked gene pairs that have been identified in *S. latifolia* is about 20% (BERGERO *et al.* 2007; FILATOV 2005; NICOLAS *et al.* 2005). The fact that the latter value is seemingly greater than the autosomal divergence from *S. conica* is most likely due to uncertainties in these estimates.

S. latifolia occurs throughout Europe and commonly grows in sunny fields and open vegetation, forming fertile hybrids with its sister species, *S. dioica*, which carries the same sex chromosome system and has a similar but more northerly distribution within Europe, and grows in more shady areas and woodlands (BAKER 1947; BAKER 1948; KARRENBERG and FAVRE 2008).

Synteny between sex-linked genes in *S. latifolia* and their homologues in *S. vulgaris* suggests that sex chromosomes in *S. latifolia* have evolved from an ordinary pair of autosomes, i.e. all of the eleven sex-linked genes described in *S. latifolia* map to a single autosome in *S. vulgaris* (BERGERO *et al*. 2007; FILATOV 2005).
The Silene system provides grounds to explore the early stages of *de novo* sex chromosome evolution, i.e. the period when recombination between the X and Y is already restricted, but degeneration of the Y may not be as advanced as in mammals or birds.

The Silene Y is the largest of n = 12 chromosomes, containing roughly 570 MB of sequence (LIU *et al.* 2004). In contrast to systems such as mammals or Drosophila, the Y chromosome is about 1.4 times larger than the X and largely euchromatic except for centromeric and subtelomeric DNA (MATSUNAGA 2006). Its mere size suggests that many (functional) genes may still be present; it also suggests that the accumulation of repetitive sequences and transposable elements may be the first process which occurs when selection against insertions is weakened (STEINEMANN and STEINEMANN 2005). However, a lot of this accumulation might have occurred in non-coding sequence, e.g. in introns, and the coding sequence may still be intact.

1.2.4 Deletion mapping

Functionally, the Silene Y has been studied using deletion mapping, i.e. chunks of the chromosome were deleted using Y-irradiation and the resulting mutant phenotype studied. Based on these studies, the chromosome can be divided into three functional regions: a female suppressor and an early anther maturation/stamen promoting locus on the p-arm, as well as a male fertility region on the q-arm (DONNISON *et al.* 1996; FARBOS *et al.* 1999; LARDON *et al.* 1999; ZLUVOVA J *et al.* 2007). Since *S. latifolia* plants lacking an X chromosome are not viable (VEUSKENS *et al.* 1992), some essential genes must have been deleted or are non-functional on the Y, suggesting that some degeneration has already occurred.

To study sex chromosomes and Y degeneration in *S. latifolia*, it is necessary to obtain and study genetical markers on the X and Y; only genes provide information about selectively disadvantageous mutations that may have become fixed on the Y, such as frameshift mutations or mutations that alter the amino acid composition of the gene product. Also expression studies comparing transcription levels from the X and Y are only possible using gene products; these are necessary, for example, to study the evolution of dosage compensation. Finding sex-linked genes in *S. latifolia*, however, is not that easy since the genome is huge (the haploid size is about 2,800 mb (LIU *et al.* 2004)), and there is no sign of it being sequenced in the near future.

1.2.5 Characteristics of known genes on the *S. latifolia* sex chromosomes

Previous studies have identified eleven sex-linked genes in *S. latifolia*.

Except for *SlssY*, all other Y-linked genes in *S. latifolia* seem to be functional: they do not contain frame-shift mutations in their coding sequences, K_A is generally lower than K_S, and all genes are expressed as mRNAs. Nevertheless, they do show some signs of degeneration: the rate of change at nonsynonymous sites, relative to changes at synonymous sites, K_A/K_S, is generally higher for Y-linked genes compared to their X-linked homologues (except for *Cyp-Y* (MARAIS *et al.* 2008)), possibly reflecting the fixation of deleterious amino acid variants. All of the six Y-linked genes tested by MARAIS *et al.* (2008) had a clear tendency to be expressed at lower levels than the X-linked genes, and some show signs of non-adaptive codon usage changes. In at least three cases, *SlCyp-Y*, *DD44-Y* and *SlY3*, transposable element sequences have accumulated in the introns of genes (BERGERO *et al.* 2008b; MARAIS *et al.* 2008). Segregating MITE elements are also found at higher frequency on the Y compared to the autosomes or the X, suggesting that selection is generally acting to reduce the number of these elements, but that selection against such insertions is reduced on the Y (BERGERO *et al.* 2008b). Similarly, the non-recombining part of the Papaya Y and the *D. miranda* neo-Y have accumulated non-coding sequences (BACHTROG *et al.* 2008; LIU *et al.* 2004), so this might be a general feature of evolving Y chromosomes.

The gene *MROS3-X* was first suggested to have a degenerated homologue on the Silene Y (GUTTMAN and CHARLESWORTH 1998); however, as *MROS3-X* belongs to a multi-gene family, it is not clear whether the degenerated copy identified on the Y is actually the original homologue of the X or a paralogue translocated from an autosome, especially since divergence between the two copies is very high, about 30%. The spermidine synthase gene, *SlssX*, is the only gene identified on the X that might have a degenerating Y-linked copy: *SlssY* has undergone several amino acid changes that presumably impair spermidine synthase activity, even though the gene is still expressed in males. Recently, FILATOV (2008) showed that *SlssX* underwent a selective sweep, possibly as a response to the degeneration of its Y-linked homologue, even though the selection coefficient associated with the sweep was presumably less than 1%. Of the eleven X-linked genes, which have been discovered until now, only *SlAp3* has no counterpart on the Y, being a duplication from an autosome onto the X (MATSUNAGA *et al.* 2003). None of the genes identified until now are directly involved in sex determination, but are classified as class I sex-linked genes, i.e. genes with presumably house-keeping functions.

1.2.6 Diversity of Y-linked genes

Theory predicts that interference among selected sites (discussed above) should lead to a reduction in diversity and adaptation of evolving Y chromosomes. Due to the size of the Silene Y chromosome, we expect many intact genes to be still present, and, with many sites under selection, we expect these interference effects to be considerable. Indeed, in Silene, all Y-linked genes that have been investigated show reduced levels of polymorphism, i.e. diversity is about 20 times lower compared to the X (MARAIS *et al.* 2008). Recent studies suggest that this is indeed due to a reduction in diversity on the Y, and not due to an unusually high diversity on the X since autosomal diversity is also high (BERGERO *et al.*, unpublished data). For older, gene-poor, Y chromosomes (Drosophila, mammals, birds), little interference among sites due to lack of recombination is expected to occur.

1.2.7 Evolutionary strata

In *S. latifolia*, recombination between X and the Y is confined to the chromosomal ends, the pseudoautosomal region (PSA), which locates to the p arm on the X and the q arm on the Y. Silent divergence data between X-and Y-linked genes suggests that recombination ceased in a gradual or step-wise manner, with genes close to the pseudoautosomal region being the least diverged (BERGERO *et al.* 2007; FILATOV 2005; NICOLAS *et al.* 2005). This is similar to the situation in birds or mammals, where genes fall into different “evolutionary strata” (LAHN and PAGE 1999; NAM and ELLEGREN 2008), though the time-scales involved in Silene are much smaller. Hence, all differences that we observe between the Silene X and Y today occurred in the relatively short time-frame after recombination between the sex chromosomes, or parts thereof, became restricted. More genes need to be studied to determine if recombination cessation in Silene has occurred in a truly step-wise manner, whether it is continuous, or whether the pattern disappears once more genes have been added. This might give insights into the mechanisms involved – which are currently

unknown - and whether it was a selective process. So far, two Y inversions have been inferred, but they seem to have occurred after recombination was already absent from the regions involved, since the inversions span genes of different X-Y divergence levels (BERGERO *et al.* 2008a). It is possible that other rearrangements on the Y or modifiers of recombination, such as DNA methylation, might have lead to suppression of recombination between the X and the Y, or that the accumulation of sexually antagonistic genes might have played a role (see chapter 4).
Clearly, more genes need to be mapped on both the X and Y to test for a causal link between gene movements, chromosomal rearrangements and recombination cessation.

1.3 Results of this study

Assuming that most new mutations that change the composition of a protein are strongly deleterious, large regions of reduced recombination are expected to show very low levels of variability. However, the observed levels of diversity on the *D. melanogaster* fourth chromosome and the *D. miranda* neo-Y chromosome do not fit the predictions of the background selection model, which predicts much lower levels of diversity for these chromosomes. In chapter 2, I describe computer simulations to show that, with very large regions of tight linkage, strongly deleterious mutations do not act independently. Instead, interference among strongly selected sites leads to their maintenance in the population for longer than predicted by the background selection model. This increases variability at linked neutral sites, to levels as observed in region of low recombination in Drosophila. Also the strong distortion in the frequency spectrum at segregating sites on the neo-Y is compatible with selection against deleterious mutations alone.

The observed high rate of accumulation of major loss-of function mutations on the *D. miranda* neo-Y is a puzzle, given that the effects on fitness associated with such mutations are expected to be large. In chapter 3, I describe computer simulations to show that the rate of degeneration of the neo-Y chromosome can be explained by interference effects among loss-of function mutations alone, without the need to invoke the action of selective sweeps. The rate of accumulation of loss-of function mutations is accelerated by the presence of mutations at nonsynonymous sites and their effects on N_e, but the impact of these mutations depends on the size of the non-recombining region and the level of Y chromosome degeneration.

In chapter 4 and 5, I describe the investigation of the evolution of sex chromosomes in *Silene latifolia*. I have isolated two new sex-linked genes in this species, *Sl-cyt* and *SlX9/Y9*, and analyzed them with respect to Y-chromosome degeneration and divergence from their X-linked chromosomal counterpart. I found evidence for the first X-linked gene in *S. latifolia* that has been translocated from an autosome to the X, and it shows signs of a selective sweep. Its possible role in having caused recombination arrest between the evolving X and Y is discussed. The second gene identified in this study, *SlX9/Y9*, still has an intact Y-linked copy that is presumably functional. Nucleotide diversity at *SlY9* is very low, whereas *SlX9* has an unusually high diversity. *SlX9* shows signs of introgression from *S. dioica* into *S. latifolia*, but the effect of this seems very localized.

2 The effects of deleterious mutations on evolution in non-recombining genomes

Contributing authors:

- I wrote the FORTRAN.95 files, performed the analysis and wrote the thesis chapter
- B. Charlesworth advised on the project, helped with the analytical methods and assisted in writing the chapter
- B. Charlesworth calculated the integral over the log-normal distribution of *s* and the confidence interval for silent diversity on the *Drosophila miranda* neo-Y chromosome

This work has been published in the following research paper:

KAISER, V. B., and B. CHARLESWORTH, 2009 The effects of deleterious mutations on evolution in non-recombining genomes. Trends in Genetics 25: 9-12

The results of this chapter made up a substantial part of this conference paper, written by B. Charlesworth:

CHARLESWORTH, B., A. J. BETANCOURT, V. B. KAISER and I. GORDO, 2009 Genetic Recombination and Molecular Evolution, in *Cold Spring Harbor Symposium on Quantitative Biology,* in press

2.1 Abstract

Under tight linkage, evolution at any one site is not independent of evolution at other sites, leading to a reduction in effective population size, N_e. There is, however, a discrepancy between the observed levels of nucleotide diversity in regions of low recombination of Drosophila, and those predicted under the background selection model: with many linked sites under selection, the reduction in N_e is consistently overestimated. To investigate if Hill-Robertson interference among strongly selected sites undermines the effects of background selection, computer simulations were carried out using parameters of mutation, selection and recombination, appropriate for deleterious amino-acid mutations in *Drosophila melanogaster* populations. The results show that genetic variability in regions of low recombination in Drosophila can be explained by interference among strongly deleterious mutations and that selection becomes progressively less effective in influencing the behaviour of neighbouring sites as the number of closely linked sites on a chromosome increases.

2.2 Introduction

When recombination rates are low, evolution does not behave according to the rules of single site models; selection acting on some sites of a chromosome will also affect measures of diversity and the efficacy of selection at linked neutral or weakly selected sites. This phenomenon, which is known as Hill-Robertson interference (HRI), is thought to play a major role in the advantage of sex and recombination (COMERON *et al.* 2008; FELSENSTEIN 1974) .

2.2.1 The background selection model

Background selection is one type of HRI, acting when selection against deleterious mutations is strong ($N_e s$ >1), so that mutations entering the population are rapidly eliminated (HALDANE 1927). Under this model, mutations are in mutation-selection equilibrium, i.e. the frequency, q_i, of a strongly deleterious mutation at the ith site in an infinite randomly mating population is solely determined by the rate of mutation from wild-type to mutant, μ_i, and the (heterozygous) selection coefficient, s_i, so that

$$q_i = \mu_i / s_i \qquad (2.1)$$

(ORR 2000)

If all sites affect fitness independently and are in linkage equilibrium, the fraction of the population that is free of deleterious mutations at all sites under selection is equal to

$$f_0 = \prod_{i=1}^{n} \left(1 - \frac{\mu_i}{s_i}\right) \qquad (2.2)$$

which can be approximated by

$$f_0 \approx exp(-(U/s)) \qquad (2.3)$$

where U is the net deleterious mutation rate over all sites ($\Sigma\mu_i$), and s is the harmonic mean selection coefficient against deleterious mutations (NORDBORG *et al.* 1996). Note that, in the corresponding diploid model, $f_0 \approx exp(-(2U/2hs))$, where U is the mutation rate per haploid genome and h is the dominance coefficient. In this classical model of background selection, only gametes falling into the mutation-free class of chromosomes will ultimately contribute to future generations because deleterious alleles are eliminated from the population with certainty. This selection regime has consequences for the fate of neutral alleles that are linked to a locus under selection because only those variants that are found on chromosomes free of deleterious mutations can survive. All others are eliminated along with the mutations found in their genomic background (hence the term "background selection") - unless they can unhitch themselves from their genomic background by recombination (HILL and ROBERTSON 1966; MCVEAN and CHARLESWORTH 2000).

2.2.2 Weak selection Hill-Robertson interference

If selection is less strong ($N_e s < 1$), deleterious mutations will spend a longer time segregating at intermediate frequencies before they get lost or - more rarely - become fixed. This is the "weak selection" Hill-Robertson effect (HILL and ROBERTSON 1966).

Under weak selection, neutral sites linked to a locus under selection are not eliminated from the population in a deterministic fashion, but variants linked to a deleterious mutation have a reduced chance of long-term survival. If two chromosomes in the population carry a beneficial variant at different selected sites, selection will increase the frequency of both haplotypes simultaneously. The mutations segregating at the two loci will impede each other's fixation and will be found less often on the same chromosome than expected from random sampling, leading to a build-up of negative linkage disequilibrium (CHARLESWORTH *et al.* 1993a). Ultimately, only one of the two advantageous variants can become fixed, unless recombination brings them together onto a single chromosome (KIMURA 1983).

2.2.3 Consequences of reduced levels of recombination

Under both background selection and "weak selection", the effective population size, N_e, of the genomic region is reduced. Under Hill-Robertson interference, any site linked to a locus under selection experiences a higher variance in reproductive success than under random sampling, i.e. the (variance) effective population size is reduced (BIRKY and WALSH 1988). In particular, under background selection, the effective population size is reduced to the fraction of the population that is free of deleterious mutations, f_0N_e (KIMURA 1983). Assuming semi-dominance such that the selection coefficient against homozygotes is $2s$, the fixation probability, u, of a selected allele at frequency p is given by the following equation:

$$u(p) = (1-e^{-Sp})/(1-e^{-S}) \qquad (2.4)$$

where $S = 4N_es$ (NORDBORG *et al.* 1996).

Accordingly, the relative strength of selection depends on the *scaled* parameter N_es, and a decrease in the effective population size is expected to lead to reduced levels of adaptation. As confirmed in computer simulations (HUDSON and KAPLAN 1995; NORDBORG *et al.* 1996), with reduced levels of recombination, the chance for a deleterious mutations to become fixed is increased, whereas the fixation probability for beneficial mutations is reduced.

An important consequence of a reduction in N_e is a reduction in neutral nucleotide diversity, π, given by the standard formula $\pi = 4N_e\mu$ (LOEWE and CHARLESWORTH 2007). The reduction in neutral variability due to background selection in the presence of recombination can be calculated using a simple equation that takes into account the effect of multiple, partially linked loci on a neutral focal site (ANDOLFATTO 2001; SHELDAHL *et al.* 2003). This assumes that each selected site acts independently with multiplicative fitness, and nonsynonymous sites are at mutation-selection equilibrium (LOEWE and CHARLESWORTH 2007).The expected level of neutral nucleotide diversity, π, relative to the level of diversity under free recombination, π_0, is estimated as

$$B = \frac{\pi}{\pi_0} \approx exp - \sum_{i=1}^{n} \frac{u_i}{s_i\left(1 + \frac{(1-s_i)r_i}{s_i}\right)^2} \qquad (2.5)$$

where u_i is the mutation rate at the ith selected site; r_i is the recombination frequency between the focal site and the selected ith site. The latter can be calculated as r = 0.5* (1-e $^{-2Z}$), where Z is the respective map distance in Morgans (Haldane mapping function), assuming no interference between crossovers.

2.2.4 How does the model fit Drosophila data?

Using estimates of the frequency of deleterious mutations and the selection coefficients acting upon them, predictions can be made about levels of neutral diversity. Drosophila polymorphism data indicate that most newly arising nonsynonymous mutations have selection coefficients that are large enough to fall into the parameter space of background selection; for regions of normal levels of recombination, the predicted value of *B* is consistent with Drosophila genomic data (BARTOLOMÉ and CHARLESWORTH 2006). However, the reduction in neutral diversity observed in regions of low recombination is not as great as expected under the background selection model. For example, the fourth chromosome of *Drosophila melanogaster* still harbours about 6% of the silent diversity compared to the autosomes (LOEWE and CHARLESWORTH 2007), whereas the background selection model predicts a relative level of diversity of about 0.1% (CHARLESWORTH *et al.*

1993a; NORDBORG *et al.* 1996). Similarly, silent diversity on the *D. miranda* neo-Y chromosome is about 1/100th of that of the neo-X (CHARLESWORTH *et al.* 1993b; PALSSON 2004) - which is much larger than expected under complete linkage (CHARLESWORTH *et al.* 1992). In other words, the BGS model greatly over-predicts the reduction in N_e caused by many linked sites under selection.

2.2.5 Can interference among strongly selected sites explain these patterns?

A possible explanation is that Hill-Robertson interference among strongly selected sites undermines the effects of background selection: if stretches of non-recombining DNA are long (and mutation rates are sufficiently high), many deleterious mutations will enter the population in each generation. Under these conditions, it will be harder for natural selection to purge the population of these mutations since a) many individuals will carry at least one mutation in the non-recombining part of the genome and b) without recombination, beneficial alleles cannot be combined, leading to an increased number of sites segregating in the population. Under these conditions, a chromosome carrying a limited number of deleterious mutations may have a selective advantage compared to the population as a whole and can survive for longer, a scenario already indicated in earlier studies (CHARLESWORTH *et al.* 1992).

To explore whether interference among strongly deleterious mutations can quantitatively explain the increase in neutral diversity relative to the expectation under the background selection model, computer simulations were carried out modelling sequence evolution in long regions with reduced recombination, using parameter estimates of mutation, selection and recombination as estimated for a typical *D. melanogaster* gene. Selection coefficients were relatively large, reflecting selection acting at non-synonymous sites. The simulations show that interference among strongly selected sites does occur if regions of reduced recombination are sufficiently long; interference leads to a relative increase in neutral diversity compared to the expectation under background selection, as observed for the *Drosophila* fourth chromosome and the *Drosophila miranda* neo-Y chromosome; modifications to the current model of background selection are thus necessary to describe the reduction in neutral diversity caused by many linked sites under strong selection.

2.3 Methods

2.3.1 The simulation model

The model consisted of a Wright-Fisher population of 1,000 haploid individuals, corresponding to a diploid population size of $N = 500$ (haploids were used to avoid the complications of extreme associative overdominance that can arise with strong selection in a small population (McVean and Charlesworth 2000). Each individual contained a single chromosome with L biallelic sites ("basepairs"), where L ranged from 3.2kb to 1.28Mb in the different runs. A chromosome was represented by a set of computer words, with the state of a given bit in a word representing the state of a nucleotide site (Loewe and Charlesworth 2007).

Two adjacent, selected ("nonsynonymous") sites alternated with one neutral ("synonymous") site along the whole length of each chromosome, roughly reflecting codon structure. Each site could be either in state "0" or "1"; for the selected sites, representing the preferred and unpreferred states respectively. Mutation, gene conversion and crossing over were simulated using bit manipulation procedures (Loewe and Charlesworth 2006), which change the state at a given position from "0" to "1" and vice versa (Figure 2.1).

2.3.2 Parameters

Rates of mutation, selection and recombination for "free recombination", multiplied by N, were chosen to match estimates of the corresponding parameter estimates from Drosophila, multiplied by the effective population size, N_e, which we set to 1.3 million (Nordborg *et al.* 1996). To a good approximation, the outcome of the evolutionary process is determined by these scaled parameters (Keightley and Eyre-Walker 2007; Loewe and Charlesworth 2006). In this way, simulations of small populations can be run that represent the behaviour of much larger populations.

The selection coefficient, s, against a deleterious mutation at a given site was drawn from a log-normal distribution with a shape and location parameter of $\sigma_g = 3.022$ and $\mu_g = 0.0368$, respectively; these correspond to the exponentials of the standard deviation and mean of ln(s) (Charlesworth *et al.* 1992). The log-normal distribution is not defined for values of $s = 0$ (i.e. neutral nonsynonymous mutations are not included in the model), so that the harmonic mean selection coefficient, s_h, can be calculated. s_h is the dominant term in a Taylor expansion of equation 2.5 for low rates of recombination, determining the reduction in diversity under background selection (HILLIKER *et al.* 1994; LOEWE and CHARLESWORTH 2007). With $N = 500$, the chosen values of σ_g and μ_g give a harmonic mean selection coefficient, such that $Ns_h = 10$. This corresponds approximately to the mean selection coefficient for mutations that are segregating in the population (LOEWE and CHARLESWORTH 2007). In our model, all sites for which $s \geq 1$ were re-assigned a selection coefficient of one ("lethal" mutations). The fraction of nonsynonymous sites where lethal mutations occurred was 0.142%, and the fraction of effectively neutral nonsynonymous mutations (for which $Ns < 1$) was less than 0.5%. Hence, the vast majority of selection coefficients lay within the range for which background selection formulae are expected to apply (MCVEAN and CHARLESWORTH 1999). This is somewhat stronger selection than indicated by analyses of *Drosophila* polymorphism data (TAJIMA 1989), so that we are probably slightly underestimating the reduction in intensity of BGS caused by HRI.

The sequence of events in each generation consisted of i) mutations entering the population, ii) selection on "adult" individuals and iii) reproduction. The number of mutations in each generation was drawn from a Poisson distribution, with an average per base pair mutation rate of $\mu = 1.04 \times 10^{-5}$ ($N_e\mu = 5.2 \times 10^{-3}$). Rates of mutation were constant along the chromosome, with the probabilities that 1 mutates to 0 and 0 to 1 being equal.

To avoid handling extremely small absolute fitness values with long chromosomes, relative fitness values were determined by a log-transformation: the log to the base e fitness, $\ln(w_i)$, of an individual who did not carry any lethal mutations was calculated as:

$$\ln(w_i) = \sum_i \ln(1 - s_i) \qquad (2.6)$$

where s_i is the selection coefficient at site i. Selection was thus multiplicative across sites. The value of $\ln(w_i)$ was compared to the expected equilibrium mean fitness of the population (SCHAEFFER 2002):

$$\ln(\tilde{w}_i) = \ln(w_i) - U \qquad (2.7)$$

where *U* is the genomic mutation rate to deleterious alleles, i.e. the total number of nonsynonymous sites times the mutation rate per site, μ. Each $\tilde{w}_i$ was divided by the maximum value of $\tilde{w}_i$ of the respective generation, to give the relative fitness of the individual. If the individual carried one or more lethal mutations, the relative fitness was set to zero. In each generation, 1,000 pairs of surviving haploid individuals were chosen (sampling with replacement), with the chance of being chosen being proportional to the relative fitness of the individual.

Three different scenarios for recombination were modelled: (a) no recombination (b) gene conversion only and (c) crossing over and gene conversion. Under scenario (a), the chromosome was treated as a single unit, i.e. the population reproduced asexually. Under scenarios (b) and (c), recombination events occurred in the diploid zygotes, and the products were used to form the pool of haploids from which the next generation was formed.

Nr_g, the scaled probability of initiation of a gene conversion event between two homologous chromosomes, was set to 9.23 x 10^{-3}, corresponding to a per base pair gene conversion frequency of 0.25 x 10^{-5} for an effective population size of 1.3 x 10^{6}. The tract length *t* was drawn from an exponential distribution with a mean of 352bp, as estimated for the *rosy* locus of *D. melanogaster* (MCVEAN and CHARLESWORTH 2000). The number of gene conversion events per generation was drawn from a Poisson distribution, and their locations placed randomly on the chromosome. Gene conversion was simulated by replacing *t*/2 bits on either side of the locus of initiation with bits from the homologous chromosome, and vice versa. If a gene conversion tract extended beyond the end of a chromosome, this end was fully converted and the tract length was accordingly shorter (see Figure 2.1 for schematic view of methods).

Under scenario (c), reciprocal crossover events between adjacent bases occurred with a constant frequency along the chromosome, where Nr_c = 0.013, corresponding to the value for regions with normal rates of crossing over in *D. melanogaster* (see (LOEWE and CHARLESWORTH 2007) for details). The number of crossover events was drawn from a Poisson distribution, and each event was simulated by exchanging the strands of the two homologous chromosomes at a random site on the chromosome.

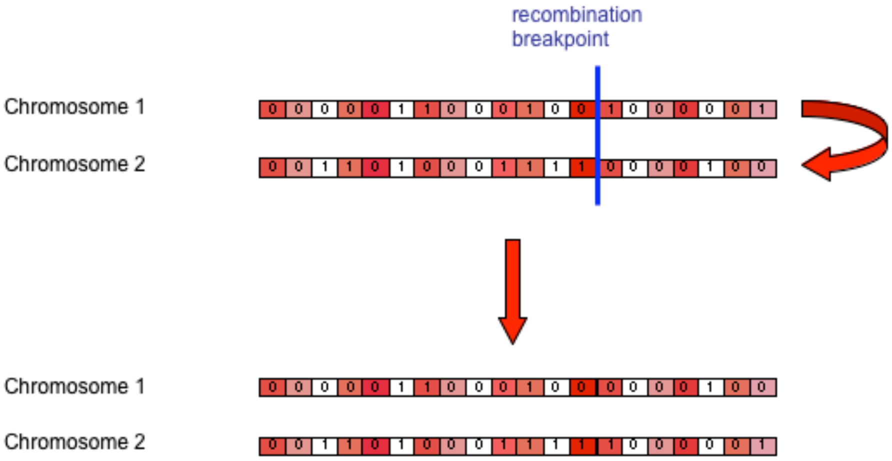


Figure 2.1: Schematic view of methods. Two selected sites (red) alternate with one neutral site (white) along each chromosome. Sites can be in state “0” or “1”. For the selected sites, the different intensities of red colour reflect possible different strengths of selection at each site. Shown here: a crossover event between two chromosomes. The breakpoint is chosen randomly along the sequence and strands are exchanged from the breakpoint onwards.

At the start of each run, the population was set to be in mutation-selection balance at each *i*th nonsynonymous site, i.e. $q_i = \mu_i/s_i$. For each nonsynonymous site on each chromosome in the population, a random number between zero and one was drawn; if this number was smaller than the value of μ_i/s_i, the particular site was set to “1”. Because mutations were assigned randomly, the population was initially at linkage equilibrium. Neutral sites were either fixed for “0” or “1”, which both occurred with a frequency of $(2 + 2N(4\mu(1+\ln[2N]))^{-1} = 0.429$ (MCVEAN and CHARLESWORTH 1999), or they were polymorphic.

2.3.3 Testing the predictions of the background selection model

Sequence evolution was simulated using a program written in FORTRAN 95 and run on the computer cluster provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk), which is partially supported by the eDIKT initiative (http://www.edikt.org). The code can be found in the appendix of this thesis.

It was confirmed that the summary statistics describing variation at both selected and neutral sites were in equilibrium after 10,000 generations of mutation, selection, and reproduction. For each parameter combination of recombination and chromosome length, four runs of 10,000 generations were performed and the average values of population statistics were calculated.

The summary statistics included Tajima's statistic, D_T, (MCVEAN and CHARLESWORTH 1999) and D_{rel} (MCVEAN and CHARLESWORTH 2000), the value of D_T relative to its maximum possible magnitude given the number of segregating sites (S). Using D_{rel} allows us to compare the bias in the frequency spectrum of polymorphic sites in the different runs, since the absolute value of D_T, is biased upwards when S is low. Other statistics that were calculated included linkage disequilibrium (D) between adjacent selected sites (MCVEAN and CHARLESWORTH 1999); the average selection coefficient at fixed nonsynonymous sites that carried the deleterious allele; neutral and nonsynonymous diversity.

Levels of neutral diversity were compared to those expected under BGS; the expected reduction in neutral variability was calculated using equation 2.5.

Each chromosome was divided into 10 bins of equal length; B(*expected*), i.e. the expected level of neutral diversity, relative to the free recombination case, was calculated for a focal neutral site in the middle of each bin (equation 2.5). This was done taking into account all nonsynonymous sites along the chromosome as well as the rates of recombination between the focal neutral site and the sites under selection; the mean of this was used for the results presented in the Figures of the Results section. For this purpose, the net map distance, z_i, taking gene conversion and reciprocal crossing over events into account, was calculated as:

$$z_i = d_i r_c + 2r_g(1 - \exp[-d_i/d_g]) \qquad (2.8)$$

where d_i is the distance between the neutral site and the selected site in terms of numbers of sites, r_c is the rate of crossing over between two adjacent bases, r_g is the probability of a gene conversion event including a particular site, and d_g is the mean tract length of a gene conversion event (2002). The observed ratio of neutral diversity relative to π_0, $B(observed) = \pi_S/(4N\mu)$, can then be compared to $B(expected)$.

A reduction in N_e is expected to have two opposing effects on nonsynonymous diversity (π_A); on the one hand, diversity is reduced as fewer individuals contribute to variability ($4N_e\mu$ is reduced). On the other hand, a reduction in the efficacy of selection (reduced N_es) increases nonsynonymous diversity because a larger fraction of deleterious mutations become effectively neutral. The approximate expected value of π_A for a given selection coefficient can be calculated using the following equation (2003):

$$\pi_A \approx \frac{2\mu(S-1)}{s(1+S)} \qquad (2.9)$$

where $S = \exp(4N_es)$.

Equation (2.9) was used to calculate the expected levels of nonsynonymous diversity, given the effective population size estimated from neutral diversity, and the integral over the log-normal distribution of s. A good match of the predicted and observed values of π_A implies that the estimate of N_e based on neutral diversity predicts the equilibrium nonsynonymous diversity.

2.4 Results

2.4.1 Test runs

It was confirmed that the average linkage disequilibrium ($D' = D/|D_{max}|$) over all adjacent segregating sites was about zero in sample runs with no selection and high recombination rates, and that measures of nucleotide diversity and Tajima's D were in equilibrium after 10,000 generations when selection operated without recombination (Figure 2.2).

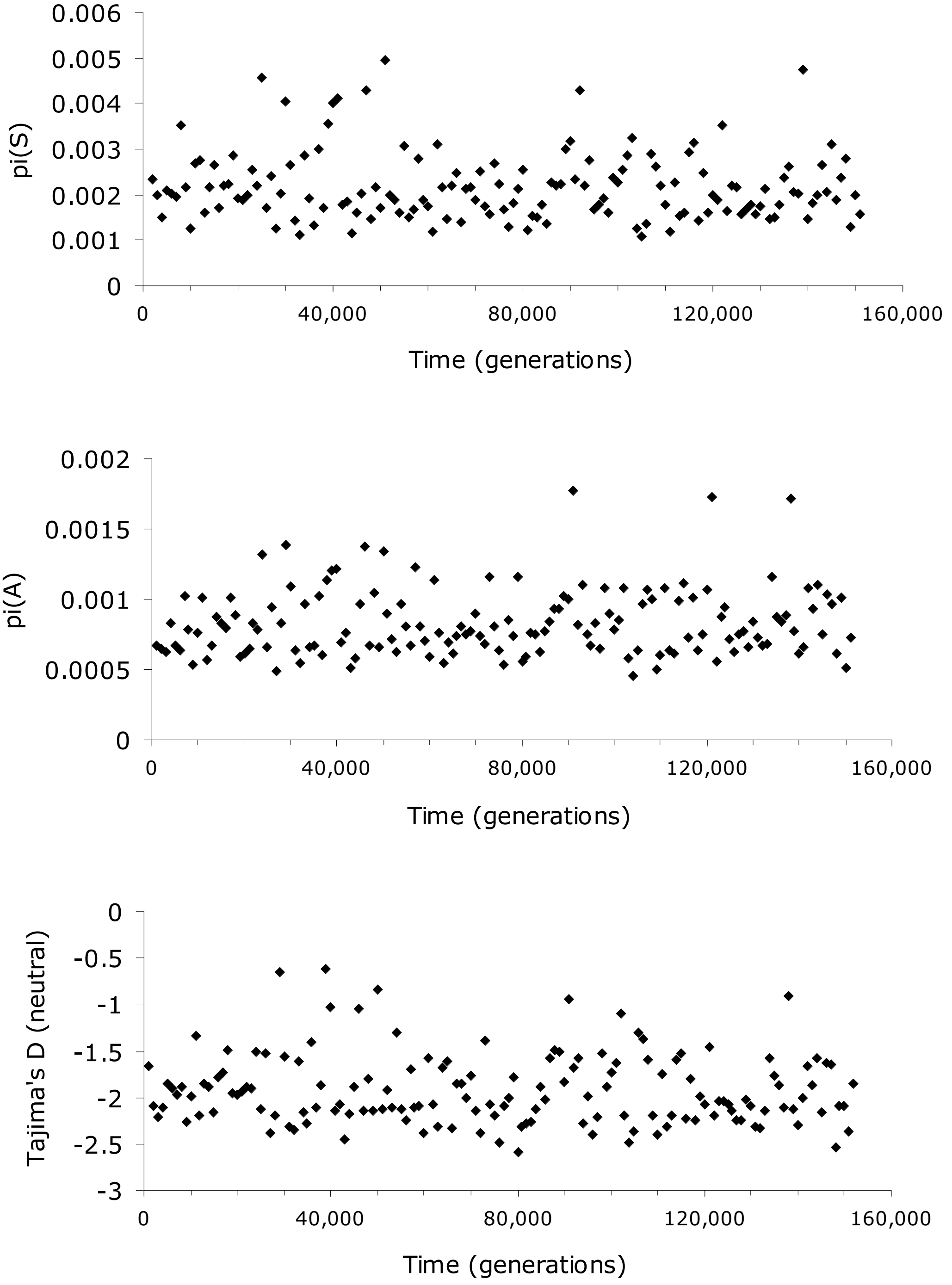


Figure 2.2: Testing for equilibrium. **(a)** Synonymous diversity, π_S **(b)** nonsynonymous diversity, π_A **(c)** Tajima's D (D_T) at synonymous sites. Statistics are plotted against the number of generations for simulations without recombination and a chromosome length of 32kb. Equilibria are reached almost immediately.

Note that, depending on the length of the chromosome, deleterious mutations could still be accumulating in the population when statistics of nucleotide diversity were sampled after 10,000 generations. The decline in population mean fitness was log-linear a long time, as expected under a Muller's Ratchet-like process of accumulation of deleterious mutations (Figure 2.3). However, an equilibrium state in fitness was reached eventually, i.e. with back mutations, fitness did not decline indefinitely.

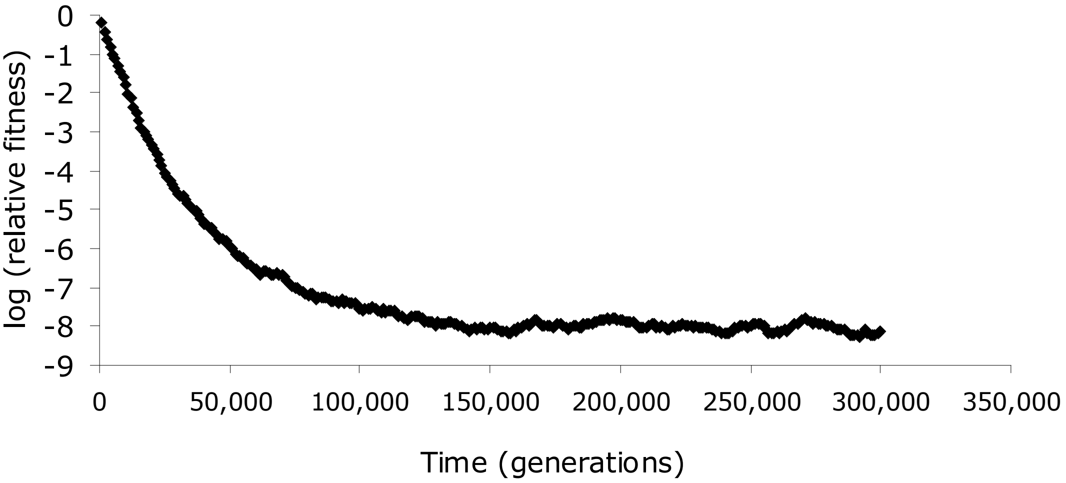


Figure 2.3: The decline in average fitness with no recombination, relative to the expected fitness under free recombination, plotted against the number of generations for simulations without recombination and a chromosome length of 32kb.

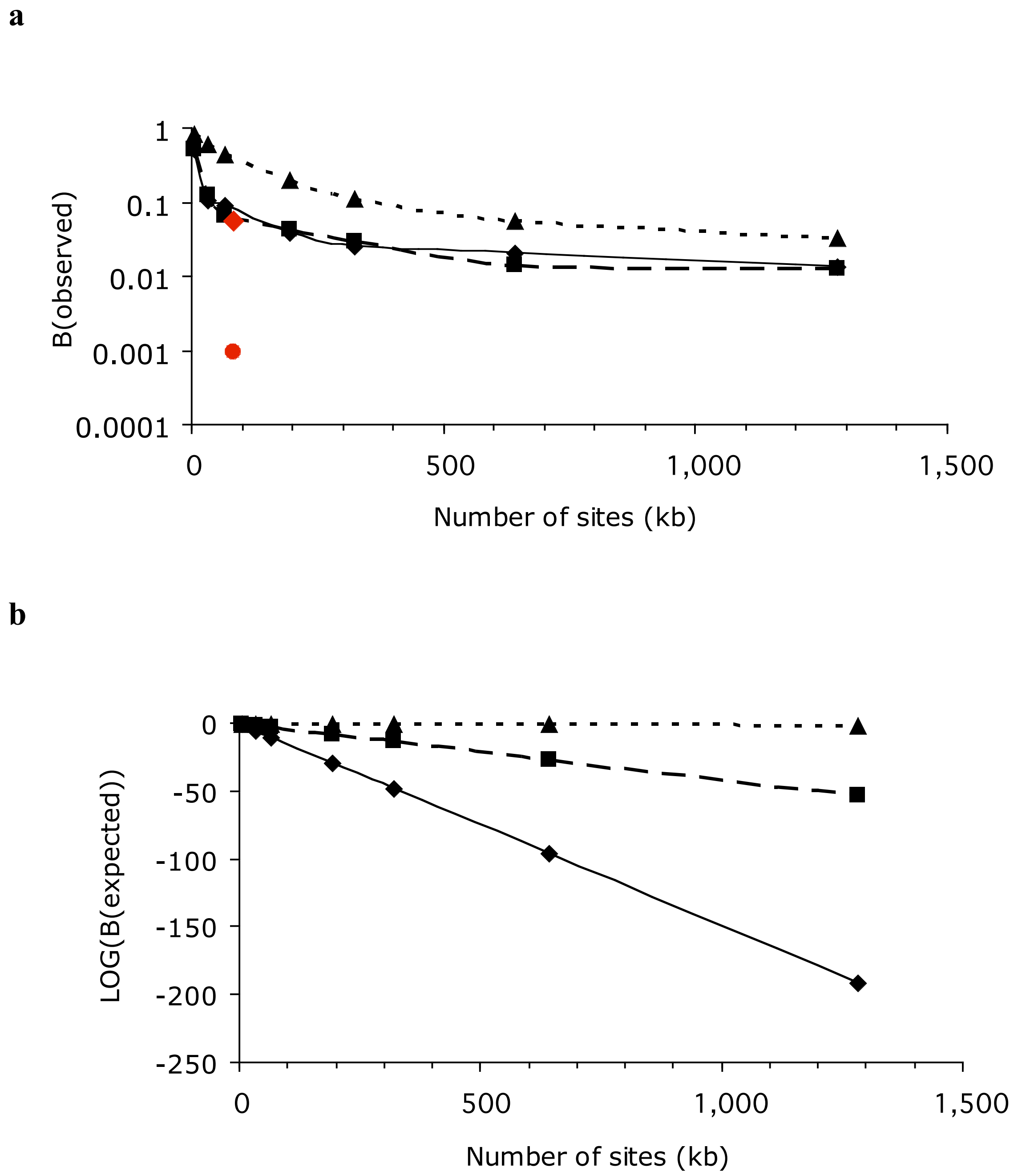
2.4.2 The effects of strong selection on neutral diversity at linked sites

When there was no crossing over, relative levels of neutral diversity, $B = \pi/\pi_0 = \pi/(4N_e\mu)$, initially decreased rapidly with an increasing number of sites, but B levelled off, reaching an asymptotic relative value of about 1.5 % for > 640,000 sites (Figure 2.4 a) To investigate how the observed reduction in diversity agreed quantitatively with the predictions under the background selection model, the expected values of B were calculated, averaging over each chromosome. With an increasing number of sites, neutral diversity is expected to decline exponentially, and this decline is more pronounced if recombination rates are low (Figure 2.4 b). Next, the observed values of B were compared with the expected reduction in diversity using equation 2.5: If both crossing over and gene conversion were allowed for, the observed strength of background selection was similar to the expected strength for all values of L. However, under the scenarios of complete linkage or gene conversion only, levels of neutral diversity increased exponentially relative to their expected values (Figure 2.4.c). In other words, with low recombination, the reduction in neutral diversity due to background selection levelled off, so that adding more selected sites to the chromosome did not reduce diversity any further.

Gene conversion generally had little effect on measures of diversity and other population statistics sampled (see below). As equation 2.8 shows, adding more sites to a chromosome does not increase the map length very much under gene conversion only; this is because the term (1 - exp[-d_i/d_g]) increases very slowly with an increase in d_i; in contrast, the map length increases linearly with L if recombination is caused by crossing overs.

2.4.3 Nonsynonymous sites under strong selection

Interference among nonsynonymous mutations also changed summary statistics on the deleterious mutations themselves. Deleterious mutations had an increased chance of fixation (Figure 2.5), and the ratio of nonsynonymous over synonymous diversity, π_A/π_S, increased with an increasing number of sites, and the effect was more pronounced if recombination rates were low, as shown in Figure 2.6. These results reflect the fact that deleterious mutations were less efficiently removed from the population when interference became more pronounced. In contrast to this, under the classical background selection model, these statistics are unaffected by the number of sites on the chromosome, and the average frequency of the unpreferred state, q, is given by equation (2.1).


a
1
0.1
0.01
0.001
0.0001
B(observed)
0
500
1,000
1,500
Number of sites (kb)
b
0
-50
-100
-150
-200
-250
LOG(B(expected))
0
500
1,000
1,500
Number of sites (kb)

c

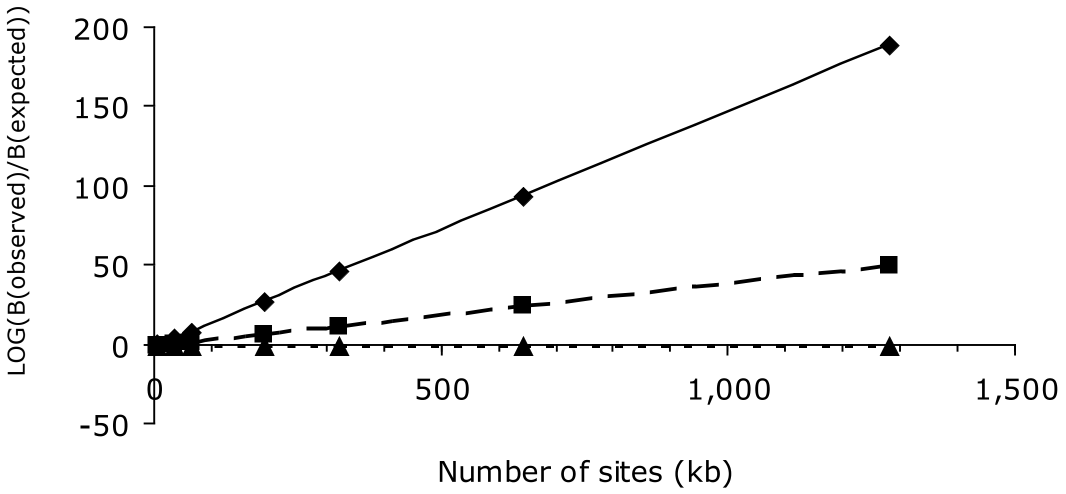


Figure 2.4: Neutral diversity is increased due to HRI among strongly selected mutations. **Symbols:** triangles, dotted line: crossing over and gene conversion; squares, dashed line: only gene conversion; diamonds, solid line: no recombination. a) Levels of neutral diversity (*π_S)*, relative to their expected value under free recombination (4*$N\mu$)*. The red circle indicates the expected levels of diversity on the *D. melanogaster* 4th chromosome under the classical BGS model, assuming independent effects of deleterious mutations, a distribution of selection coefficients and gene conversion as the only mechanism of genetic exchange (Loewe and Charlesworth 2007); the red diamond indicates the observed levels of diversity on the *D. melanogaster* 4th chromosome; these two values are plotted at 82kb, the approximate length of coding sequence on the 4th chromosome (see Discussion) **b)** The expected reduction in neutral diversity. *B(expected)* was calculated using equation 2.5. **c)** The observed reduction in neutral diversity (*B(observed)* = $\pi_S/(4N\mu)$), relative to *B(expected)*. Under low recombination, *B(observed)* / *B(expected)* increases exponentially. If recombination rates are sufficiently high, the expected and observed values of *B* correspond reasonably well.

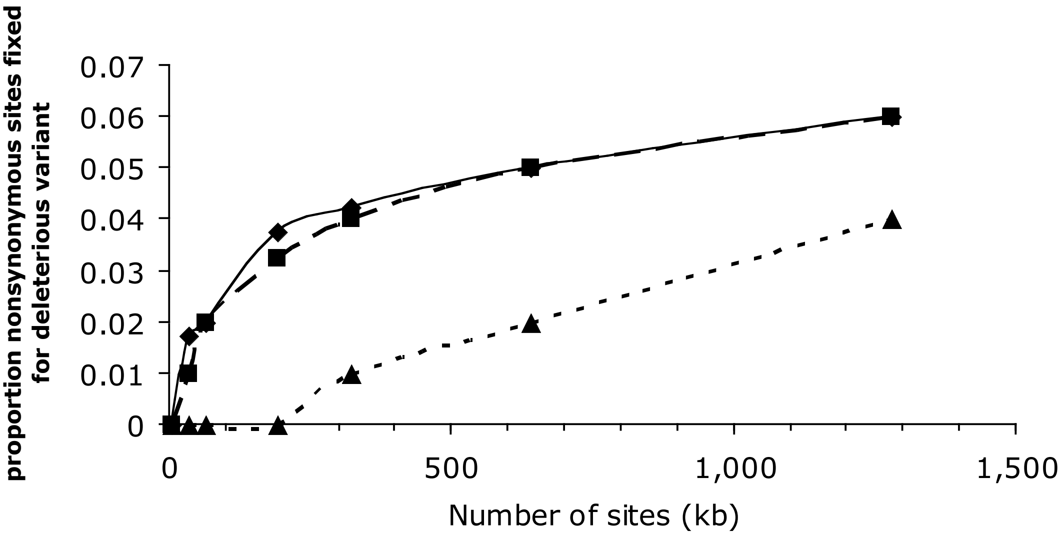


Figure 2.5: The proportion of nonsynonymous sites fixed for the deleterious variant after 10,000 generations. **Symbols:** triangles, dotted line: crossing over and gene conversion; squares, dashed line: only gene conversion; diamonds, solid line: no recombination.

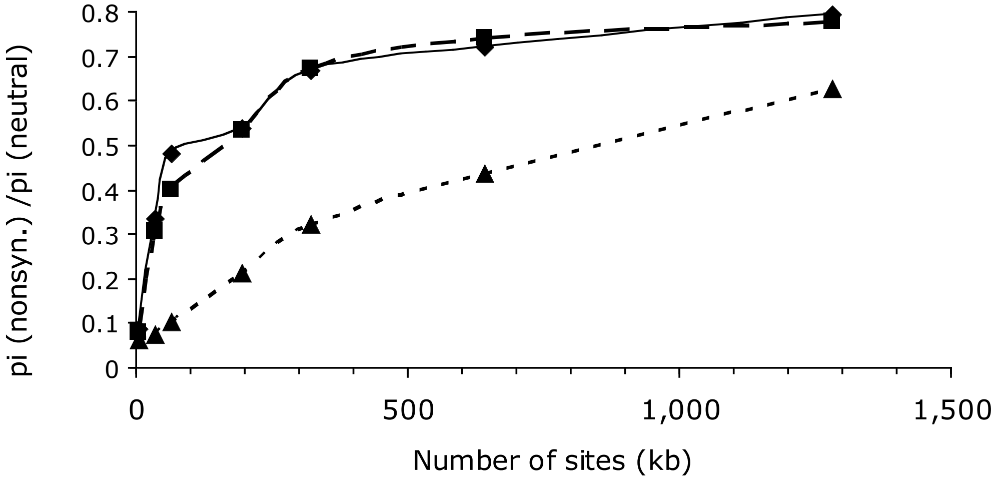


Figure 2.6: Nonsynonymous diversity relative to neutral diversity. π_A over π_S is plotted against the number of sites. The ratio increases with L as selection against deleterious mutations becomes less efficient. The increase is highest with lower recombination and is apparently reaching an asymptotic value. **Symbols:** triangles, dotted line: crossing over and gene conversion; squares, dashed line: only gene conversion; diamonds, solid line: no recombination.

Nonsynonymous diversity decreases as interference becomes stronger because the absolute number of variants maintained in the population decreases with decreasing N_e (2007). Estimates of nonsynonymous diversity agree very well with their predicted levels based on equation (2.9), given the effective population size estimated from synonymous diversity and the integral over the distribution of selection coefficients (Figures 2.7 a) and 2.7 b)). Hence, estimates of N_e based on either π_A or π_S are very similar; both statistics indicate that N_e is larger than expected under the background selection model if many selected sites are relatively closely linked. Note that, since the observed values of π_A are averages of only four simulation runs, fluctuations in Figure 2.7a are most likely due to random chance; the apparent increase in diversity at 64kb for the no recombination case is due to an outlier of $\pi_A = 0.00133$, giving a standard error of 30% for this data point.

2.4.4 Distorted genealogies, linkage disequilibrium and reduced efficacy of selection

Selection distorts gene genealogies as shown by a negative Tajima's *D* (Figure 2.8 a) that was generally lower when recombination was reduced or absent. Hence, interference does not only lead to a simple reduction in the effective population size, but it also has an impact on the frequency distribution of segregating sites. For the neutral sites, D_{rel} values were decreased to nearly minus one with an increasing number of sites, suggesting that most variants were in fact singletons at this stage (Figure 2.8 b)

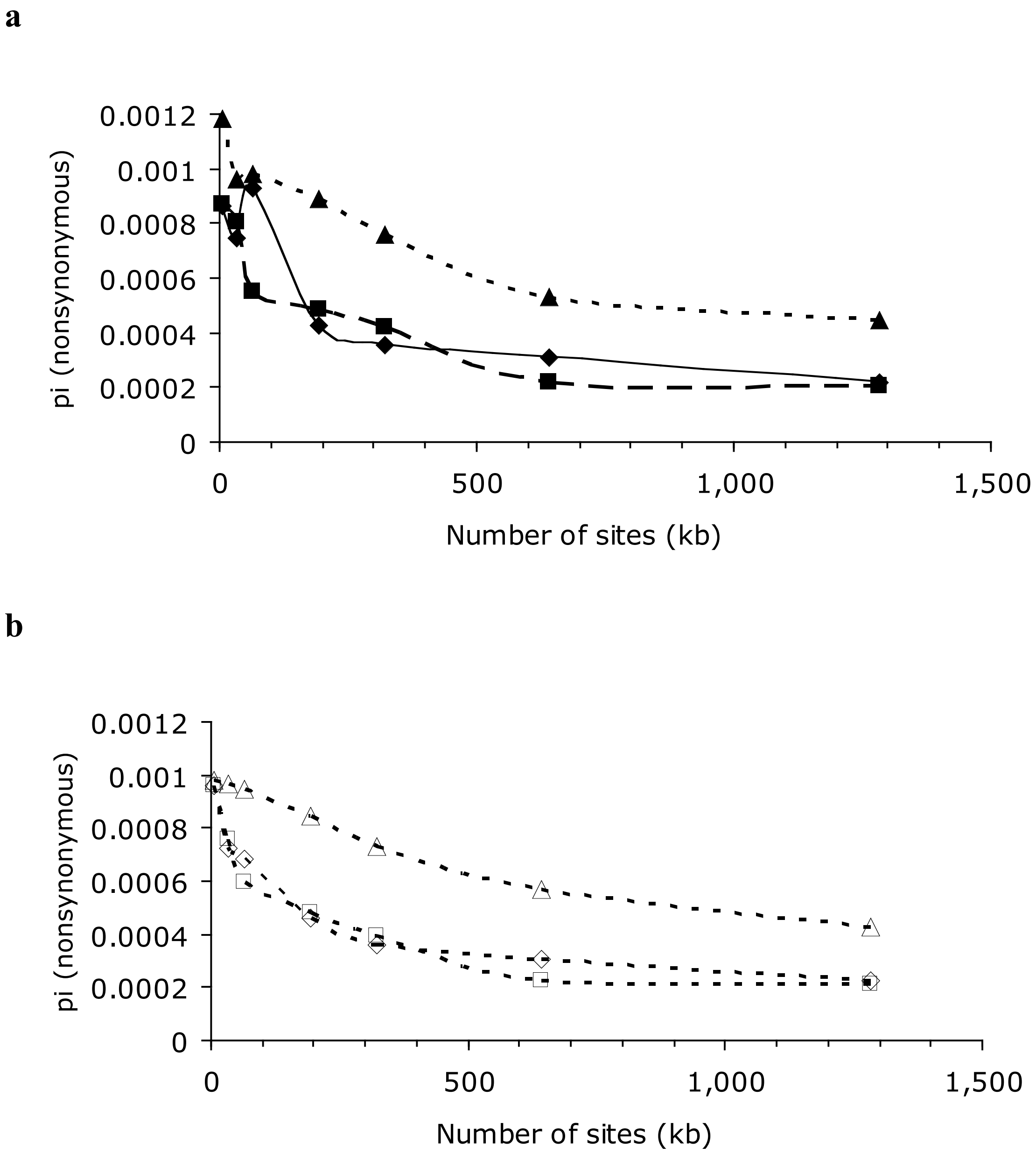


Figure 2.7: a) Nonsynonymous diversity, π_A, as observed in the simulations. **Symbols:** triangles, dotted line: crossing over and gene conversion; squares, dashed line: only gene conversion; diamonds, solid line: no recombination. **b)** The expected levels of π_A calculated from equation (2.9). **Symbols**: Open triangles: crossing over and gene conversion; open squares: only gene conversion; open diamonds: no recombination.

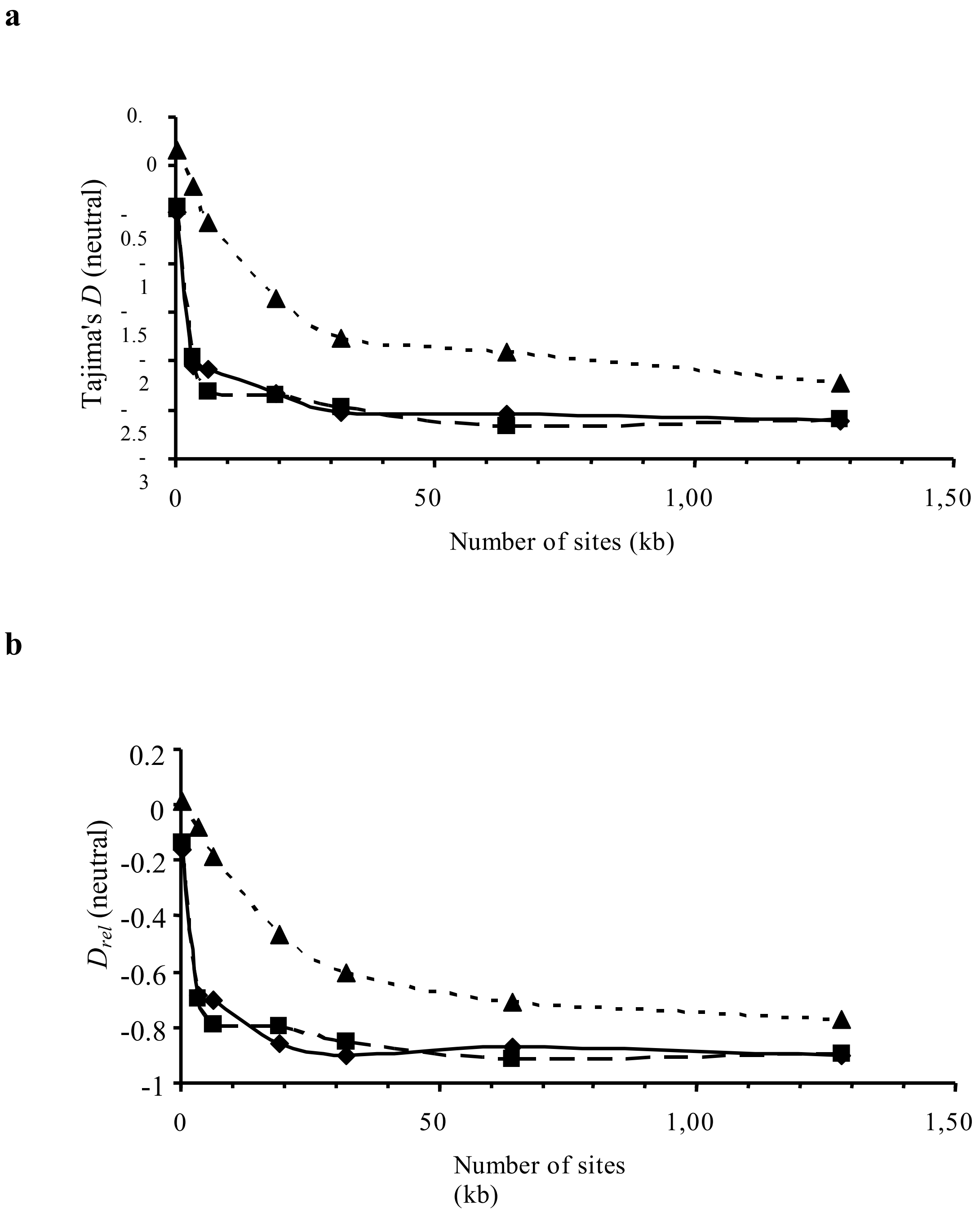


Figure 2.8: The skew in the frequency distribution of polymorphic sites. a) Tajima's *D* values and b) D_{rel} for neutral sites. **Symbols:** triangles, dotted line: crossing over and gene conversion; squares, dashed line: only gene conversion; diamonds, solid line: no recombination.

As shown in Figure 2.9, the average selection coefficient at nonsynonymous sites that were fixed for the unpreferred state increased with an increasing number of sites on the chromosome: with very short chromosomes (3.2 kb) and no recombination, the arithmetic average $N_e s$ for fixed deleterious mutations was about 1; for 1.28Mb of completely linked sites, this value increased to about 18, reflecting the fact that selection against deleterious mutations was less efficient with increasing interference.

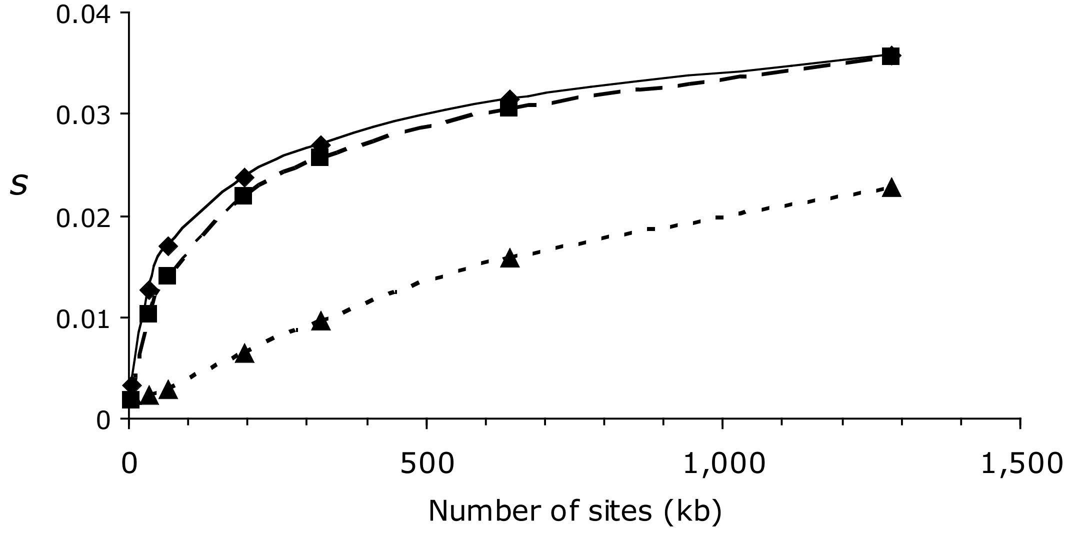


Figure 2.9: The average selection coefficient at nonsynonymous sites that were fixed for the deleterious variant, plotted against the number of sites on the chromosome. **Symbols:** triangles, dotted line: crossing over and gene conversion; squares, dashed line: only gene conversion; diamonds, solid line: no recombination.

Linkage disequilibrium (*D'*) between adjacent selected segregating sites was always negative (Figure 2.10), suggesting the existence of repulsion haplotypes – a general characteristic of Hill-Robertson interference; this effect did not seem to increase with an increasing number of sites.

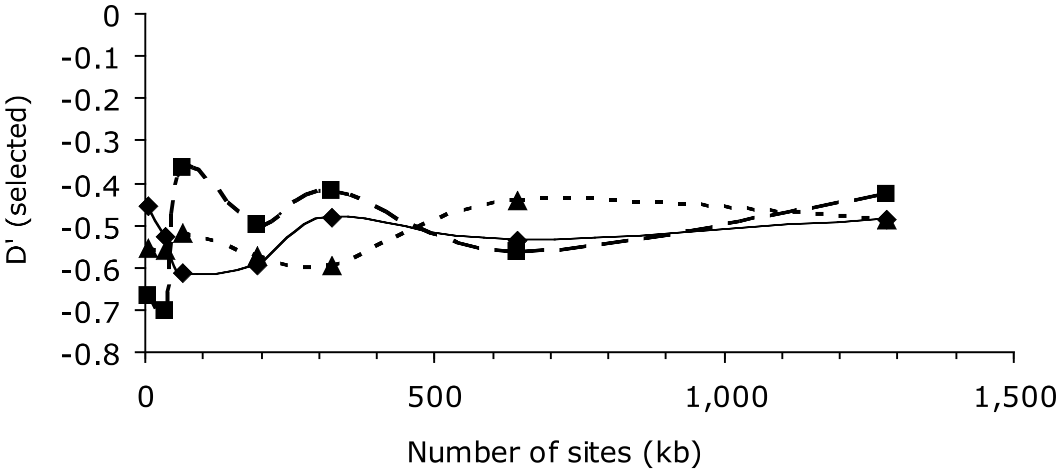


Figure 2.10: Linkage disequilibrium (measured as *D´* (2006)) between adjacent selected segregating sites, plotted against the number of sites on the chromosome. **Symbols:** triangles, dotted line: crossing over and gene conversion; squares, dashed line: only gene conversion; diamonds, solid line: no recombination.

There was no evidence that sites in the middle of the chromosome experienced stronger effects of background selection that those at the ends (results not shown), as expected from equation 2.5. However, given the large number of sites on each chromosome, the effect that was expected was only very weak: the maximum difference for the reduction of neutral diversity of sites in the centre of the chromosome versus those at the ends was expected for the shortest chromosomes (3.2 kb) undergoing crossing over and gene conversion. But even for this parameter combination, the difference between *B(predicted)* in the middle vs. the ends of the chromosome was only 3%; for 320kb, the expected difference was less than one percent.

In contrast to the simulations with reduced recombination, in runs that included both crossing over and gene conversion at normal rates, the observed reduction in neutral diversity was actually greater than expected under the background selection model (Figure 2.11); this effect is not eminent in Figure 2.4c where the data are plotted on a log-scale.

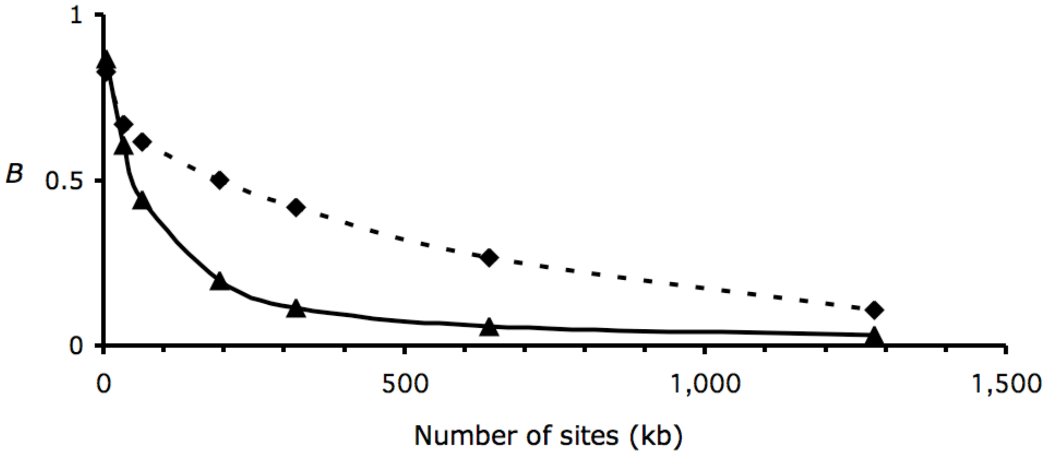


Figure 2.11: The observed reduction in neutral diversity (triangles, solid line) and the expected values of *B* under the background selection model (diamonds, dotted line), plotted against the number of sites on the chromosome.

2.5 Discussion

The simulations show that the predictions of the background selection model break down if many strongly selected sites are relatively tightly linked. With a reduction in N_e, deleterious variants behave as if they were subject to weaker selection and are removed from the population less efficiently. This maintains them in the population for longer and increases variability at linked neutral sites. In other words, interference among *segregating* selected variants is not confined to weakly selected sites - such as synonymous sites that are selected for optimal codon usage (2006)- but can also occur among amino acid sites.

How do our results relate to observations from genomic regions with low recombination? The fourth chromosome of *D. melanogaster* contains roughly 85 genes; it does not cross over under normal conditions, although it can experience gene conversion (BACHTROG *et al.* 2008). Assuming an average length of 965 coding sites per gene (as given for 82 genes at http://flybase.bio.indiana.edu/), the chromosome contains roughly 82,000 partially linked coding sites. For chromosome 4, mean values of π_{silent} of 1.28 x 10^{-3} for 18 loci and 1.32 x 10^{-3} for four loci were obtained for worldwide collections by WANG *et al.* and

SHELDAHL (1974), giving an overall mean of approximately 1.3 x 10^{-3}. The data of SHAPIRO *et al.* (LOEWE and CHARLESWORTH 2007) on autosomal variation for genes with normal levels of recombination gave a mean of 0.023 for silent sites. The ratio of π values for chromosome 4 to genes with normal recombination is thus about 0.06. This level of diversity is very similar to the value obtained in the simulations, but much larger than predicted under the background selection model (Figure 2.4).

The neo-Y chromosome of *Drosophila miranda* lacks recombination, and contains roughly 3.7 Mb of coding sequence (3087 genes of about 1,200 sites), although about half of the genes have lost their function and are thus unlikely to cause HRI. Even though neutral diversity on the neo-Y chromosome is strongly reduced, it is still detectable; BARTOLOMÉ and CHARLESWORTH (1996) observed an average silent diversity of 4 x 10^{-5} for 20 genes on the *D. miranda* neo-Y. The confidence interval on this value can be estimated approximately as follows. The high degree of distortion of the genealogy of this chromosome (CHARLESWORTH 1996; LOEWE and CHARLESWORTH 2007) means that the standard equilibrium coalescent approach cannot be used. Instead, we assume that the genealogy is a star phylogeny, which is close to what is suggested by the data. A minimum width on the confidence interval of the estimated nucleotide site diversity, π, is then provided by a Poisson distribution of the number of segregating sites in the sample (this ignores any stochastic variation in the depth of the genealogy). If all variants are singletons, π is estimated as $2S/(nm)$, where S is the number of segregating sites, n is the number of alleles in the sample, and m is the number of sites sequenced (for any given site, $n - 1$ of the alleles in the sample differ from the allele with the variant, and there are $n(n - 1)/2$ pairs of alleles). Pooling all the sites in the data of BARTOLOMÉ and CHARLESWORTH , we have $S = 9$, $n = 18$ and $m= 23{,}064$, giving an estimate of 4.35 x 10^{-5} for π. This is very close to the π value obtained directly (there is no difference between diversities for nonsynonymous and silent sites, reflecting the lack of effective selection on the nonsynonymous sites). The lower and upper 2.5 percentile values for the mean of the Poisson from the Poisson distribution for 9 observed events are 4.78 and 17.1, respectively, corresponding to π values of 2.30 x 10^{-5} and 8.26 x 10^{-5}. The corresponding 95% confidence interval ratio of the neo-Y π to the mean silent site π (3.91 x 10^{-3}) for the same genes on the neo-X is 0.0059 to 0.021. Given that there are three times as many neo-

X as neo-Y chromosomes, the corresponding interval for the ratio of effective population sizes is 0.018 and 0.063. In practice, the interval must be substantially wider than this, since the value for the neo-Y is broader than our estimate, and we have ignored the error for the neo-X (assuming independence among loci, the mean for the neo-X has an s.e. of 1.0 x 10^{-4} .
In the simulations, relative levels of diversity for complete linkage relative to the expected value under free recombination were about 0.015 for the longest chromosomes simulated (*L* = 1.28 Mb); these sequences were not much shorter than the total length of coding sequence that is presumably still functional on the neo-Y . We conclude that HRI among amino acid sites can explain diversity on regions of low recombination in Drosophila, such as the fourth chromosome or the neo-Y.

BACHTROG (2004) argued that the reduction in diversity on the *D.miranda* neo-Y chromosome was most likely to be explained by positive, rather than negative selection, based on coalescent simulations, which suggested that the observed Tajima's *D* value of about -2 was more likely to occur under a selective sweep scenario rather than background selection. However, her BGS simulations assumed independent effects among sites and did not take interference into account. In our forward simulations, on the other hand, frequency spectra of segregating sites were very strongly skewed towards low frequency variants, giving Tajima's *D* values as low as -2.5. This suggests that selection against deleterious mutation may indeed result in Tajima's *D* values as low as observed for the neo-Y chromosome.

With a harmonic mean $N_e s$ of 10, interference can occur even with realistic levels of recombination as estimated for *D. melanogaster* autosomes: as shown in Figure 2.5, the proportion of sites fixed for the deleterious allele also increased with increasing chromosome length in simulations that had crossing over and gene conversion. This might imply that a substantial number of deleterious amino acid mutations could become fixed also in regions of normal recombination, i.e. a substantial fraction of nonsynonymous substitutions between species could be deleterious. MARUYAMA and KIMURA (1974) showed that the time to fixation, conditional on fixation, is the same for a deleterious and a positively selected mutation that have selection coefficients of equal magnitude.

Accordingly, deleterious mutations that are destined to go to fixation will leave traces in adjacent genomic regions resembling selective sweeps, and this phenomenon may be more common that previously thought. Note, however, that we did not allow for spacing between genes, whereas the *D. melanogaster* genome has an average spacing of 6kb; this is likely to decrease interference effects among selected sites in regions of normal levels of recombination .

With recombination, background selection was more efficient than predicted by the model (Figure 2.11), similar to the results of NORDBORG *et al.* (1996), who simulated much shorter chromosomes and observed less marked effects. This might be a result of simulating small populations in which negative linkage disequilibrium builds up among deleterious mutations at closely linked sites, as shown in Figure 2.10. Equation (2.5) assumes independence between sites under selection, and this assumption is clearly often violated. Negative linkage disequilibrium implies that the total frequency of haplotypes carrying at least one deleterious mutation is higher than with linkage equilibrium, for the same allele frequencies at each site. If selection against deleterious mutations is sufficiently strong that a neutral variant associated with a single closely-linked deleterious mutation has a high chance of elimination before it recombines away, this means that the efficacy of BGS will be increased. The true value of *B* for regions of normal recombination on a Drosophila chromosome is likely to be close to 1, and the discrepancy with the simulations is possibly caused by the lack of intergenic sequence simulated.

3 The rate of gene loss on the *Drosophila miranda* neo-Y chromosome can be explained by the process of Muller's ratchet

Contributing Authors:

- I wrote the FORTRAN.95 files, performed the analysis and wrote the manuscript
- B. Charlesworth advised on the project, helped with the analytical methods and assisted in writing the manuscript

3.1 Abstract

Since its formation about 1.75MYA, the *Drosophila miranda* neo-Y chromosome has undergone a rapid process of degeneration, having lost approximately half of the genes that it originally contained. Using estimates of mutation rates and selection coefficients against loss-of-function mutations, I show that the high rate of accumulation of these mutations can be explained by the process of Muller's ratchet, the stochastic loss of the least-loaded mutational class from a small asexual population. I show that selection at nonsynonymous coding sites can accelerate the process of gene loss, and that this effect varies with the number of genes still present on the degenerating neo-Y chromosome.

3.2 Introduction

Without recombination, sites in the genome do not evolve independently of each other (FELSENSTEIN 1974; FISHER 1930; MULLER 1932), which leads to reduced levels of nucleotide diversity and adaptation, such as non-optimal codon usage or a high rate of amino acid changes (BACHTROG 2003; 2005; BARTOLOMÉ and CHARLESWORTH 2006; BETANCOURT and PRESGRAVES 2002; BETANCOURT *et al.* 2009; CHARLESWORTH and CHARLESWORTH 2000). It is, however, still an open question how large-scale re-arrangements and the loss of whole open reading frames can become fixed in non-recombining regions of the genome, leading to structures such as the small and degenerate Y chromosome of humans or the W chromosome in chicken (FRIDOLFSSON *et al.* 1998; SKALETSKY *et al.* 2003).

The neo-Y chromosome of *D. miranda* is an example of a large non-recombining region that is relatively young and has only partially degenerated (BACHTROG *et al.* 2008; BARTOLOMÉ and CHARLESWORTH 2006; STEINEMANN and STEINEMANN 1998), enabling us to study the time-frame over which degeneration can occur - as well as its possible causes. The neo-Y arose when an autosome (corresponding to chromosome arm 2R in *D. melanogaster*) became fused to the Y chromosome, containing about 3,000 genes with a total of about 3.7 Mb coding sequence (STEINEMANN and STEINEMANN 1998). Since there is no recombinational exchange between homologues in Drosophila males (GETHMANN 1988), recombination between the neo-X and neo-Y became immediately restricted; within a short evolutionary time-frame of only ~ 1.75 MY (BARTOLOMÉ and CHARLESWORTH 2006), about half of the genes originally present on the neo-Y have lost their function (BACHTROG *et al.* 2008).

In order to quantify the rate of accumulation of loss-of-function mutations on the neo-Y chromosome, which we will denote by r, it is convenient to consider the base-line rate of fixation for neutral mutations (which is equal to the mutation rate), and compare this to the observed rate of fixation of "major" mutations. BACHTROG *et al.* (2008) showed that 55/118 genes present on the ancestral neo-Y contain at least one frame-shift mutation, stop codons, or deletion, while these genes have remained intact in the neo-X lineage. Accordingly, with an average length of neo-Y linked coding sequence of 1188bp in this dataset (BACHTROG *et al.* 2008), the divergence per basepair with respect to loss-of-function mutations (K_D) is given by $K_D = (55/118)/1188 = 3.9 \times 10^{-4}$. These mutations have all occurred along the neo-Y branch of the tree connecting the neo-Y and neo-X chromosomes to their common ancestor. The corresponding synonymous site divergence, K_S, is about 1% (BARTOLOMÉ and CHARLESWORTH 2006). To estimate the neutral level of divergence along this branch with respect to indel mutations – which probably are the main cause of loss of gene function (see below) - we multiply the value of K_S by 0.44, i.e. the amount of new mutations that cause indels in Drosophila, relative to those causing transitions and transversions (HAAG-LIAUTARD *et al.* 2007). If U is the rate of origination of major deleterious mutations on the neo-Y, the data thus suggest that r/U on the neo-Y, given by $K_D/0.44K_S$, is about 9%, a rather high value. This approach has the advantage of being independent of the number of generations per year in *D. miranda*, although it might lead to an underestimate of r/U since it ignores the possibility of multiple loss-of function mutations within the same open reading frame (see Discussion).

We do not know what processes have driven the rapid accumulation of loss-of function mutations. We expect the selection coefficients associated with the heterozygous carriers of such mutations to be rather large (CROW and SIMMONS 1983), and deleterious mutations for which $N_e s >> 1$ (where N_e is the effective population size and s is the selection coefficient) have a very low probability of fixation (KIMURA 1983). Accordingly, one or more forces must be acting to severely reduce the N_e of the neo-Y. Positive selection, causing selective sweeps (KAPLAN *et al.* 1989; MAYNARD SMITH and HAIGH 1974), can drag to fixation deleterious, linked variants, provided that selection at the beneficial sites is strong enough to overcome the cumulative effect of selection against deleterious mutations in the background (CHARLESWORTH 1994; HADANY and FELDMAN 2005; JOHNSON and BARTON 2002). However, an unrealistically high incidence of strong positive selection is probably necessary to explain the neo-Y data on this basis (see Discussion).

In this chapter, I therefore examine an alternative "null" model that does not invoke selective sweeps. Assuming that deletions, frameshift mutations and insertions of transposable elements are irreversible, I will examine the process of Muller's ratchet as a means for fixing major mutational lesions (FELSENSTEIN 1974; HAIGH 1978; MULLER 1964), as previously proposed for the evolution of Y chromosomes by CHARLESWORTH (1978). Under this model, selection against deleterious mutations is sufficiently strong ($N_e s > 1$) so that mutations in the freely recombining, ancestral population are close to mutation-selection equilibrium. With multiplicative fitness effects and a Poisson distribution of the number of mutations per haploid genome, the equilibrium size of the mutation-free class in a Wright-Fisher population is given by $N_0 = N \exp(-U/s)$, where N is the population size in terms of number of haploid genomes, U the genomic mutation rate for deleterious mutations for the chromosome in question, and s the selection coefficient (HAIGH 1978). If the population size is finite, genetic drift will eventually lead to the stochastic loss of this class of individuals; without recombination, it cannot be restored (the ratchet has made one "click"). The process of repeated loss of the least-loaded class of individuals leads to the constant accumulation of deleterious mutations within the population, and with each "click" of the ratchet, one deleterious mutation becomes fixed (CHARLESWORTH and CHARLESWORTH 1997). The rate of fixation of deleterious mutations, r, is thus greatly increased over that for mutations with the same selection coefficients in a freely recombining population.

We can reasonably assume that major mutations are irreversible and that $N_e s$ for such mutations is much larger than one. However, we do not know a priori whether the ratchet can explain the neo-Y data, since it cannot operate if $N_0 s$ is too large (GORDO and CHARLESWORTH 2000b). One factor that might speed up the ratchet is the presence of deleterious mutations caused by base substitutions at amino-acid sites in coding sequences. We will call these sites "background selection" or "BGS" sites. As shown in chapter 2, selection at BGS sites can drastically reduce the N_e value for a non-recombining genomic region, but the reduction in N_e levels off as the number of nonsynonymous sites under selection increases. For very long chromosomes such as the neo-Y, neutral diversity, which is directly proportional to N_e (KIMURA 1983), asymptotes at a level of about 1.5% of the value with free recombination. We expect the reduction in N_e caused by the BGS sites to accelerate the rate of the ratchet, since GORDO and CHARLESWORTH (2001) have shown that background selection can have such an effect, but the expected magnitude of the effect is unknown for realistic parameters.

In this chapter, I show that a high rate of fixation of strongly deleterious loss-of-function mutations on the neo-Y chromosome of *D. miranda* is compatible with a "null" model of selection acting against deleterious mutations alone. I also show that selection against amino acid mutations, which are under weaker selection than loss-of-function mutations, has a significant effect on the rate at which major mutations can accumulate.

3.3 Methods

3.3.1 Theoretical background

Analytical and numerical results are available for the speed of the ratchet in a non-recombining, Wright-Fisher haploid population of size N, where the rate of origin of new deleterious mutations per generation is U, the selection coefficient against a single mutation is s, and fitness effects of different mutations combine multiplicatively (GESSLER 1995; GORDO and CHARLESWORTH 2000a; b; 2001; HAIGH 1978; HIGGS and WOODCOCK 1995; JAIN 2008; PAMILO *et al.* 1987; ROUZINE *et al.* 2008; STEPHAN *et al.* 1993).

We will make use of some of these results for interpreting the rate of movement of the ratchet. We first consider the case when $N_0 > 1$, where $N_0 = N\exp(-U/s)$ – see above. Following a click of the ratchet, the population will approach a new equilibrium after time, T_A, with the number of individuals carrying just one mutation being equal to $N\exp(-U/s)$. T_A has been estimated (GORDO and CHARLESWORTH 2000a) as

$$T_A \approx \frac{1}{s}\left(1 - \frac{1.6s}{U}\right) \qquad (1)$$

Recently, JAIN (2008) has derived an analytic expression for the average time, T_C, between two clicks of the ratchet (disregarding T_A), provided that $N_0 >> 1$:

$$\overline{T}_C \approx \begin{cases} 2N_0\left[1 + ln\left(1 + \frac{1}{\sqrt{\beta}}\right)\right], & \beta << 1 \quad (2a) \\ \sqrt{\pi}N_0\beta^{-3/2}e^{\beta}, & \beta >> 1 \quad (2b) \end{cases}$$

where $\beta = cN_0s$. As pointed out by JAIN (2008), with $c = 0.6$, the integral used to derive equations (2) is identical to the one described in GORDO and CHARLESWORTH (2000b), which is why we will use $c = 0.6$ here.

The net expected rate of fixation of deleterious mutations, r, is thus given by $1/(T_A + T_C)$. This can be compared with the value for a freely recombining haploid population of size N

$$r = \frac{2UNs}{e^{2Ns} - 1} \qquad (3)$$

(KIMURA 1962).

The significance of N_0 and N_0s for driving the ratchet has been previously discussed (BELL 1988; GESSLER 1995; GORDO and CHARLESWORTH 2000b; HAIGH 1978; STEPHAN *et al.* 1993). Equations (2) imply that the rate of the ratchet scales linearly with N. This means that r/U, (i.e. the rate relative to the neutral rate) is expected to be constant if the products Ns and NU are held constant, as would be expected from the fact that these results are derived from a diffusion equation approximation (EWENS 2004).

In some cases that we studied, the condition $N_0 > 1$ is violated. We then used numerical solutions of equations (1) – (5) of GESSLER (1995) to compute r.

3.3.2 The model

To simulate the accumulation of major deleterious mutations by the ratchet in the presence of more weakly selected deleterious mutations, we used forward simulations of sequence evolution, similar to those described in chapter 2. Briefly, we used a Wright-Fisher model consisting of a population of 1,000 haploid individuals, each of which carry a single non-recombining chromosome of length L, where L varies from 32kb to 1.28Mb. Two-thirds of all the sites on a chromosome are "background selection sites" (BGS sites), representing sites at which non-synonymous mutations can occur. The selection coefficients for these sites are drawn from a log-normal distribution with a harmonic mean Ns for the corresponding diploid population of size 500 (i.e. the coalescent N (CHARLESWORTH 2009; HUDSON 1990)) equal to 10. At the remaining sites on the chromosome, major knock-out mutations can occur that have very large fitness effects on the individual. These sites will be called "major" sites. Selection is multiplicative across all sites. Note that, since we do not allow recombination, the exact position of the "major" sites on our simulated chromosomes is irrelevant.

At the start of each run, all BGS sites are in mutation-selection equilibrium, whereas there are no mutations at the major sites. The rate of mutations at BGS sites is constant per site, i.e. adding more sites to the chromosome increases the chromosome-wide mutation rate for the BGS sites. In contrast, we keep the chromosome-wide mutation rate, U, at the major sites constant, i.e. we measure the

effect of increasing or decreasing BGS, without changing the influx of major mutations. Mutations at major sites are irreversible and are thus expected to accumulate via a Muller's ratchet-type process; the BGS sites are reversible, and initially accumulate at a constant rate, until an equilibrium between forward and backward mutation is reached (chapter 2). The reduction in N_e caused by the BGS sites, however, reaches a steady state almost immediately (chapter 2).
10,000 generations of mutation, selection and reproduction were performed, and the rate, r, of fixation of major mutations was estimated by calculating the slope of the regression line for the number of sites fixed against time.

3.3.3 Tests of equations (1) and (2)

In test-runs, I checked whether I obtained similar rates of r/U for the "major" mutations from the simulations as suggested by equations (1) and (2), using a range of parameters for U and s. To calculate the expected rate, I used equation (2a) or (2b) respectively, depending on the value of β, and added the term T_A (equation 1) to obtain the total expected time between clicks of the ratchet. I performed simulations with or without selection at the BGS sites, as shown in Table 3-1.

Whenever we allowed BGS to occur at nonsynonymous sites, I calculated the reduction in N_e caused by the BGS sites alone, calculated from the formula for expected neutral diversity, $\pi = 4N_e\mu$, as obtained from the simulations described in chapter 2. The expected rate of accumulation of major mutations, r/U, was then calculated by replacing the term N in equations (2) and the equations of GESSLER (1995) with the estimate of N_e.

3.3.4 Mutational parameters and scaling by population size

In order to be able to compare our simulation results with the r/U value of 0.04 that we estimated for the *D. miranda* neo-Y chromosome, I have used scaled values of mutation and selection parameters. According to diffusion theory, it is possible to infer the behaviour of the system in a much larger population than assumed in the simulations by keeping the products of N_eU and N_es constant (EWENS 2004; HILL and ROBERTSON 1966; MCVEAN and CHARLESWORTH 2000), if time is measured in units of N_e generations. Without interference effects, the effective population size for the

D. miranda neo-Y is likely to be one quarter of the diploid N_e, which has been estimated to be about 840,000 (LOEWE *et al.* 2006). Mutation and selection parameters for use in our simulations with a haploid number of 1,000 are thus obtained by multiplying the biologically realistic values by 4.2 x 10^5/1,000 = 420.

We will assume here that the main two causes of loss of function of a gene are insertion-deletion (indel) mutations or TE insertions. Both types of mutations are likely to contribute equally to fitness loss, and are hence treated as a single process from the point of view of the ratchet. The rate of origination of indel mutations in *D. melanogaster* is about 2.6 x 10^{-9} per bp per generation (HAAG-LIAUTARD *et al.* 2007). With an estimated 3.67 million coding sequence sites on the neo-Y before degeneration (the size of the homologous region in *D. pseudoobscura*), this gives a per chromosome mutation rate of approximately 0.0095. We further assume that the observed frequency of TE insertions into intronic sequences in the *D. miranda* neo-Y chromosome reflects the insertion rate into coding sequences, without selective constraints. No TEs were found within coding sequence by BACHTROG *et al.*(2008), but 13 out of 118 genes that were present on the ancestral neo-Y carry new TE insertions in introns. The total length of intron sequence in the sample of genes studied by BACHTROG *et al.* (2008) is about 70.2kb, so that the number of putatively neutral TE insertions per bp is 13/(70.2 x 10^3) = i.e. 1.85 x 10^{-4}. The predicted rate of accumulation of neutral indels per bp on the neo-Y branch is about 44% of the value for base substitutions (HAAG-LIAUTARD *et al.* 2007), i.e. 0.01/2.3 = 0.0044, so that we estimate that the rate of insertion of new TEs into coding sequence relative to the rate for indels is 1.85 x 10^{-4}/0.0044 = 0.042, which can be neglected. We will use a slightly conservative estimate of the rate of "major" mutations on the neo-Y in *D. miranda*, before degeneration of U = 0.009. If we scale this value to a haploid population size of 1,000, keeping NU constant, this gives a U value of 3.78 for our simulations.

Note that the point mutation rate of $N\mu$ = 0.0052 at the BGS sites that we used in our simulations was obtained by combining the above estimate of the mutation rate per basepair with the *D. melanogaster* estimate of N_e, which is about 1.3 million (LOEWE and CHARLESWORTH 2007; LOEWE *et al.* 2006); this enabled us to estimate the reduction in N_e caused by the BGS sites alone, since data were available from previous simulations (chapter 2). A realistic value for *D. miranda* would be about one-quarter less than this, if current estimates of neutral diversity for *D. miranda* are used (LOEWE *et al.* 2006). It seems likely, however, that *D. miranda* has undergone a recent reduction in effective population size (BACHTROG 2007; BACHTROG and ANDOLFATTO 2006; YI *et al.* 2003), so that use of this larger value is probably realistic as far as the history of the neo-Y chromosome is concerned. In any case, having an increased μ per BGS site is roughly equivalent to having more sites in the simulations and hence should not affect the results substantially, especially since the reduction in N_e due to background selection levels off as the number of BGS sites increases.

3.3.5 Estimates of the selection coefficients against major mutations

About one-quarter of loss of function mutations in Drosophila are lethal in the homozygous state, but lead to a mean reduction in fitness of only 1-2% when heterozygous (CHARLESWORTH and CHARLESWORTH 1998; CHARLESWORTH and HUGHES 1999; CROW and SIMMONS 1983). Since mutations on the neo-Y are nearly always masked by functional alleles on the neo-X, we need here to consider only the heterozygous selection coefficients. Knock-out mutations are all not expected to have the same effects (i.e. losing a gene that is part of a gene family might be less deleterious than losing a single-copy gene). However, the arithmetic mean selection coefficient against mutations that are segregating in a randomly mating population can be used as an estimate for the harmonic mean, s_h, of a distribution of s values (LOEWE *et al.* 2006), because segregating mutations tend to be less deleterious than the average of all new mutations. We can assume that this distribution of selective effects does not include $s = 0$ because gene loss is unlikely to be completely neutral; hence, s_h is always defined. The heterozygous selection coefficient of segregating knock-out mutations for enzyme loci in *D. melanogaster* has been estimated to be about $s = 0.0015$ (LANGLEY *et al.* 1981). For a haploid population of $N = 1{,}000$, this corresponds to a scaled s value of 0.63.

To circumvent the problem of using a wide distribution of scaled s values, which can generate unrealistically large heterozygous selection coefficients (>> 1), we tested whether simulations with a single selection coefficient for all major sites gives quantitatively similar results compared to using a distribution of *s* values, provided that the fixed *s* value is equal to the harmonic mean of the distribution of *s* values. Runs were performed for 10,000 generations, and *r/U* was calculated and compared between runs.

3.3.6 Tests of scaling

Scaling of parameters of mutation and selection by the population size produces coherent results when the *s* values are relatively small (McVEAN and CHARLESWORTH 2000), but the diffusion approximations might break down for stronger selection, such as the parameter space assumed for selection at the major sites. We therefore tested whether the scaling of *s* by the population size works, i.e. we tested whether *r/U* is constant if the population size is changed, and parameters of mutations and selection are scaled appropriately. Runs using a population size of *N* = 1,000, 10,000, 20,000 or 40,000 individuals were performed. In these runs, all BGS sites were assigned a fixed selection coefficient that corresponds to a (diploid) *Ns* of 10, and the (diploid) *Ns* at the major sites was equal to 315. (This corresponds to a heterozygous *s* value of 0.63 for a haploid population size of *N* = 1,000, i.e. the scaled neo-Y value). We measured *r/U* for the major mutations, using a chromosome length of 32kb, with four runs performed per population size.

3.3.7 Testing how *r/U* behaves when the length of the chromosome increases

Runs using scaled neo-Y parameters of mutation and selection were performed, with different lengths of chromosomes (32 kb to 1.28 Mb), and the average *r/U* was measured for each run. Four runs were performed for each parameter combination. To compare these results to the expected rates, assuming that *N* is reduced to the N_e suggested by the BGS simulations of chapter 2, we cannot use equations (2) because the equilibrium size of the least- loaded class, $N_0 = N\exp(-U/s)$ is less than 1; in other words, the least-loaded class is never present (GESSLER 1995; JAIN 2008; ROUZINE *et al.* 2008). Hence, we used the approach of GESSLER (1995) to calculate the expected rates whenever $N_0 < 1$.

3.4 Results

The question that we wish to explore is whether the estimated value for the *D. miranda* neo-Y chromosome of r/U, the rate of fixation of major mutations relative to their mutation rate, can be accounted for by Muller's ratchet, using the parameter values and simulation methods described above. We first examine how well the theoretical formulae and rescaling by population size perform, and then describe the major results of interest for interpreting the *D. miranda* results. Results shown in Figures 3-1 and 3-4 are averages of 4 runs; all other results are from single simulation runs.

3.4.1 Tests of the theoretical formulae

Table 3-1 presents the results of testing the theoretical predictions describe above, using a single selection coefficient s for the major mutations. It can be seen that equations (1) and (2) often tend to overestimate r, sometimes quite badly, but can be used as a rough guideline for the expected rate of the ratchet when there is no selection at the BGS sites (upper part of Table 3-1). The first two rows of Table 3-1 show parameter combinations used by GORDO and CHARLESWORTH (2000b), who obtained rates that were very similar to ours, suggesting that our simulations produce comparable results ($r/U \approx 0.131$ and $r/U \approx 0.035$ from Figure 1 in GORDO and CHARLESWORTH (2000b), compared to our r/U-values of 0.129 and 0.040 respectively).

With selection at the BGS sites, we observe a large increase in the rate of the ratchet (lower part of Table 3-1); the observed rate is now substantially larger than expected from equations (1) and (2) when ignoring BGS, and the effect is more pronounced when selection at the "major" sites is not very strong ($s = 0.2$). However, the observed rate is more similar to the expected rate when the latter is calculated using the "background selection N_e" instead of N in the theoretical predictions, at least for the first two parameter combinations shown in Table 3-1, for which $s = 0.2$.

3.4.2 Tests of the scaling and selection parameters used in the simulations

Using the approach described in the Methods section, the procedure of scaling the mutation and selection parameters by the population size *N* produced very similar results for simulated values of *r*/*U*. Differences in *N* had almost no effect when the products of *Ns* and *Nμ* were held constant for both the BGS sites and the "major" sites (Figure 3-1). Hence, we can be reasonably confident that our simulations of small populations can be used to predict the behaviour of the ratchet in a much larger population, such as that of *D. miranda.*

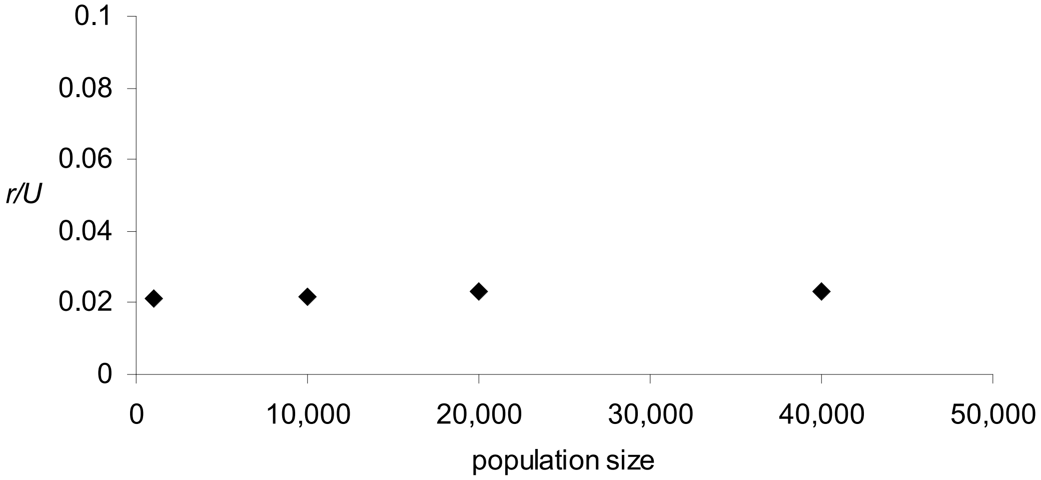


Figure 3-1: *r/U* scales with the population size *N.* At the "major" sites, *Ns* (diploids) = 315, as estimated for "major" mutations occurring on the *D. miranda* neo-Y. At the BGS sites, selection occurs with a constant selection coefficient, so that *Ns* of the corresponding diploid population = 10. The mutation rates at all sites are also scaled by the population size. *L* = 32 kb in all cases.

We also found that *r/U* remains largely unchanged if a fixed *s* value is used, instead of a log-normal distribution of *s* values, provided that the fixed *s* value corresponds to the harmonic mean of the respective distribution (Table 3-2). Hence, we can reasonably assume that our simulations using scaled values of *U* with a fixed selection coefficient produce results comparable to those of a much larger population where *s* values are drawn from a distribution of values.

Table 3-1: Testing equation 3, with corrections for T_A (equation 1). The first part of the Table gives results for simulation runs without selection at the BGS sites, i.e. considering Muller's ratchet only. The second part shows results with selection at the BGS sites combined with Muller's ratchet. In the last column, the expected rate, *r*(exp, with BGS), was calculated assuming that *N* is reduced to the N_e suggested by the BGS simulations (chapter 2).

1.)No Background selection

U	***s***	***U/s***	N_0	β	***r/U* (observed)**	***r* (obs)/*r* (exp)**	***r*(obs)/*r*(exp, with BGS)**
0.0240	0.015	1.60	202	1.82	0.040	0.86	N.A.
0.0585	0.015	3.90	20	0.18	0.129	0.97	N.A.
0.1200	0.05	2.40	91	2.72	0.005	0.36	N.A.
0.0220	0.01	2.20	111	0.66	0.080	0.75	N.A.
0.0010	0.001	1.00	368	0.22	0.436	0.69	N.A.
0.1000	0.02	5.00	7	0.08	0.129	0.88	N.A.
0.025	0.0063	4.00	18	0.07	0.197	0.90	N.A.
0.69	0.20	3.45	32	3.81	0.002	0.48	N.A.
3.78	0.63	6.0	2.5	0.94	0.021	0.78	N.A.
2.) Background selection at nonsynonymous sites							
0.69[1]	0.20	3.45	32	3.81	0.071	16.79	0.61
0.69[2]	0.20	3.45	32	3.81	0.297	70.57	1.28
3.78[2]	0.63	6.0	2.5	0.94	0.032	1.14	0.19

[1] BGS with L = 32 kb; N_e (BGS simulations) = 108

[2] BGS with L = 320 kb; N_e (BGS simulations) = 2

Table 3-2: *r/U* is given for a range of *s*-values at the "major" sites. In all runs, BGS occurs at the first and second codon position, with *Ns(harmonic)* = 10 at the BGS sites (*L* = 320 kb; U = 3.78)

1.) Distribution of *s*-values at "major" sites. *s* is drawn from a log-normal distribution with mean μ_g and standard deviation σ_g. The fraction of lethal mutations, as well as the arithmetic and harmonic mean *s* is given, considering either only non-lethal mutations or all mutations at "major" sites. The fraction of "major" sites where *s* takes values in the interval 0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8 and 0.8-1.0 is indicated.

2.) All "major" sites are assigned a fixed selection coefficient *s*. This *s*-value equals the harmonic mean *s* of a distribution in part 1.) of the Table. The *r/U*-values of part 1.) and part 2.) of the Table are very similar.

1.) *s* from a log-normal distribution:

	μ_g = -1.0 **σ_g = 0.4**	**μ_g = -1.0** **σ_g = 0.1**	**μ_g = -0.7** **σ_g = 0.4**	**μ_g = -0.7** **σ_g = 0.1**	**μ_g = -0.35** **σ_g = 0.2**	**μ_g = -0.35** **σ_g = 0.1**	**μ_g = -0.2** **σ_g = 0.1**
fraction (lethal mutations)	0.005	0.000	0.040	0.000	0.040	0.0002	0.022
s (arithmetic) non-lethal	0.394	0.370	0.511	0.499	0.704	0.708	0.818
s (harmonic) nonlethal	0.338	0.366	0.447	0.494	0.681	0.701	0.811
s (arithmetic) total	0.398	0.370	0.531	0.499	0.716	0.708	0.822
***s (harmonic)* total**	**0.339**	**0.366**	**0.457**	**0.494**	**0.689**	**0.701**	**0.814**
fraction (0 ≤ *s* ≤ 0.2)	0.064	0.000	0.011	<0.001	<0.001	<0.001	<0.001
fraction (0.2 < *s* ≤ 0.4)	0.518	0.796	0.285	0.015	0.002	<0.001	<0.001
fraction (0.4 < *s* ≤ 0.6)	0.306	0.204	0.385	0.956	0.211	0.054	0.001
fraction (0.6 < *s* ≤ 0.8)	0.086	0.000	0.201	0.029	0.521	0.843	0.408
fraction (0.8 < *s* ≤ 1.0)	0.026	0.000	0.118	0.000	0.265	0.103	0.591

Table 3-2, continued

r/U	**0.144**	**0.135**	**0.089**	**0.077**	**0.018**	**0.015**	**0.001**
2.) *s* fixed :							
s	**0.339**	**0.366**	**0.457**	**0.494**	**0.689**	**0.701**	**0.814**
r/U	**0.149**	**0.135**	**0.091**	**0.077**	**0.017**	**0.014**	**0.001**

3.4.3 The effects of BGS on the speed of the ratchet

Our simulations show that selection at the BGS sites increases the rate of fixation at major sites, but the magnitude of the effect depends on the relative strengths of selection at the BGS and major sites, as well as on the number of BGS sites. As shown in Figure 3-2, r/U is consistently higher when there are more sites under background selection (L = 1.28 Mb *versus* L = 320 kb). Here, the expected values for r/U were calculated taking the reduction in N_e due to the BGS sites into account; both the observed and the expected rate decrease with increasing s (because more strongly deleterious mutations accumulate more slowly), but the observed rate decreases faster than the expected rate. Accordingly, when s at the "major" sites – and hence the difference between the two types of selection coefficients – becomes larger, the effect of the BGS sites becomes increasingly less pronounced. The expected r/U values calculated *without* the reduction in N_e caused by the BGS sites are about two orders of magnitude lower than the observed r/U values in Figure 3-2 (Table 3-3).

Longer chromosomes (more BGS) lead to higher rates of fixation at the "major" sites (Figure 3-3). Note that, if the parameter β (a major determinant of the rate of the ratchet in the absence of BGS (GORDO and CHARLESWORTH 2000b; JAIN 2008)) is held constant, the ratio r/U decreases with increasing U/s, whereas the actual rate, r, increases (this is simply due to the non-linearity of U, s, β and the fact that U increases faster than r) (Figure 3-3). For the smallest s value at the "major sites" simulated (s = 0.06), the increase in r/U with increasing L is most pronounced (r/U increases from 3% to 78%), suggesting that BGS can have a very large effect on r, provided that s values at the two types of sites are similar.

Figure 3-4 shows the effect of selection at BGS sites on the rate of fixation of major mutations, with parameters estimated as realistic for the neo-Y chromosome of *D. miranda*, and scaled appropriately by the population size (N = 1,000; s = 0.63; U = 3.78). As in Figure 3, the rate r/U increases with the amount of background selection at nonsynonymous sites. Note, however, that for these parameter combinations, the background selection effect is not very large: even when there is no selection at BGS sites (dotted line), r/U is about 2%. The ratio r/U increases to

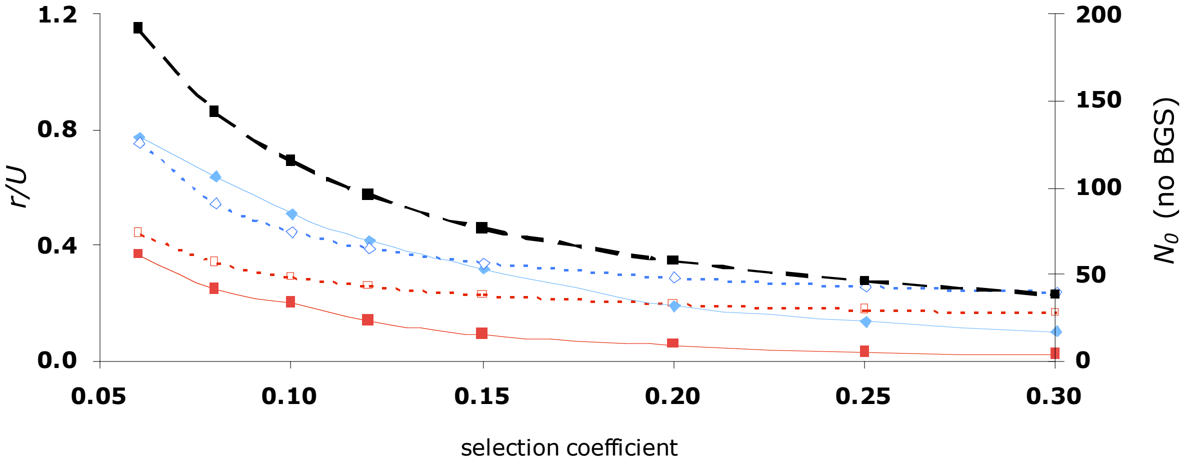


Figure 3-2: The effect of varying *s* at the "major sites", with different levels of background selection, i.e. different values of *L*. *β* (not taking the BGS effect into account) is held constant at 6.9. Since $\beta = N\exp(-U/s)0.6s$, *U* also varies with *s*; values of *U* are listed in Table 3-3. The plot shows the observed and expected *r/U* for two lengths of chromosomes (*L* = 320 kb and *L* = 1.28 Mb respectively). The expected *r/U* values were calculated assuming that *N* in equation 3 is reduced to the N_e suggested by the BGS simulations described in chapter 2. The corresponding expected *r/U* values without the BGS effect are listed in Table 3-3.

Symbols: Left-hand axis: red squares (filled symbols), solid line: observed *r/U* for *L* = 320 kb; red squares (open symbols), dotted line: expected *r/U* for *L* = 320 kb; blue diamonds (filled symbols), solid line: observed *r/U* for *L* = 1,280 kb; blue diamonds (open symbols), dotted line: expected *r/U* for *L* = 1,280 kb.

Right-hand axis: black squares, dashed line: N_0 (without the BGS effects)

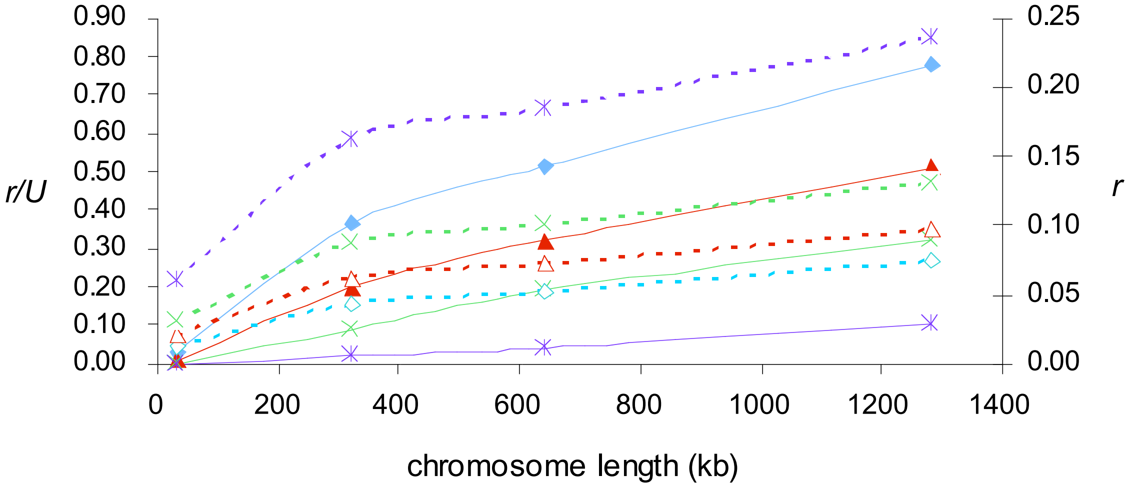


Figure 3-3: Dependence of *r/U* on *U/s* and *L*. As in Figures 1 and 2, *β* (not taking the BGS effect into account) is held constant at 6.9; **Symbols:** filled symbols, solid lines: *r* (right-hand Y axis); open symbols, dotted lines: *r/U* (left-hand axis). Blue diamonds: *U* = 0.099, *s* = 0.06, *U/s* = 1.65; red triangles: *U* = 0.216, *s* = 0.10, *U/s* = 2.16; green crosses: *U* = 0.385, *s* = 0.15, *U/s* = 2.57; purple asterisks: *U* = 0.978, *s* = 0.30, *U/s* = 3.26.

Table 3-3: Expected rates of the ratchet for “major” mutations for a range of *U* and *s* values when β is held constant. Due to the non-linearity of *U, s* and β, with an increase in s, *U* has to increases more, and the ratio *U/s* becomes larger, and hence N_0 becomes smaller. The expected rate increases with decreasing N_0, but *r/U* actually decreases due to the relatively higher increase in *U*.

U						***r* (expected, no BGS)**	***r/U* (expected, no BGS)**
	s	***U/s***	***N***	**β (noBGS)**	**N_0 (no BGS)**		
0.099	0.06	1.65	1,000	6.9	192	5.38×10^{-05}	5.42×10^{-04}
0.155	0.08	1.94	1,000	6.9	144	7.17×10^{-05}	4.62×10^{-04}
0.216	0.10	2.16	1,000	6.9	115	8.96×10^{-05}	4.14×10^{-04}
0.281	0.12	2.35	1,000	6.9	96	1.08×10^{-04}	3.82×10^{-04}
0.385	0.15	2.57	1,000	6.9	77	1.34×10^{-04}	3.49×10^{-04}
0.571	0.20	2.86	1,000	6.9	58	1.79×10^{-04}	3.14×10^{-04}
0.770	0.25	3.08	1,000	6.9	46	2.24×10^{-04}	2.91×10^{-04}
0.978	0.30	3.26	1,000	6.9	38	2.69×10^{-04}	2.75×10^{-04}

Table 3-4: The expected rates for simulations using scaled parameters as estimated for the *D. miranda* neo-Y, calculated using the approach of Gessler (1995). The expected rates based on equation 3 (with the correction of T_A) are also shown for comparison. The corresponding rates observed in the simulations are shown in Figure 4.

L (kb)		32	192	320	640	1280
$U = 3.78$ & $s = 0.63$	N_0	0.27	0.10	0.06	0.05	0.03
	r/U (Gessler)	0.00	0.17	0.17	0.50	0.33
	r/U equation 3 with T_A correction	0.11	0.15	0.17	0.18	0.19
$U = 2.5$ & $s = 0.63$	N_0	2.04	0.76	0.49	0.40	0.26
	r/U (Gessler)	N.A.	0.17	0.17	0.17	0.33
	r/U equation 3 with T_A correction	0.05	0.10	0.13	0.15	0.18
$U = 1.89$ & $s = 0.63$	N_0	5.38	1.99	1.29	1.05	0.67
	r/U (Gessler)	N.A.	N.A.	N.A.	N.A.	0.17
	r/U equation 3 with T_A correction	0.03	0.07	0.09	0.11	0.15

about twice this value for the longest chromosomes simulated (*L* = 1.28 Mb) and could be even higher for chromosomes of the size of the ancestral neo-Y chromosome (over twice as much coding sequence). Hence, the *r/U* obtained from the simulations with the value of *U*/*s* that we have proposed as plausible is similar to the value for the *D. miranda* neo-Y (*r/U* of about 9%). Table 3-4 shows the expected rates for the neo-Y data, taking the BGS effect into account. The expected rates are clearly higher than the observed rates, i.e. the N_e driving the ratchet is much larger than suggested by neutral diversity.

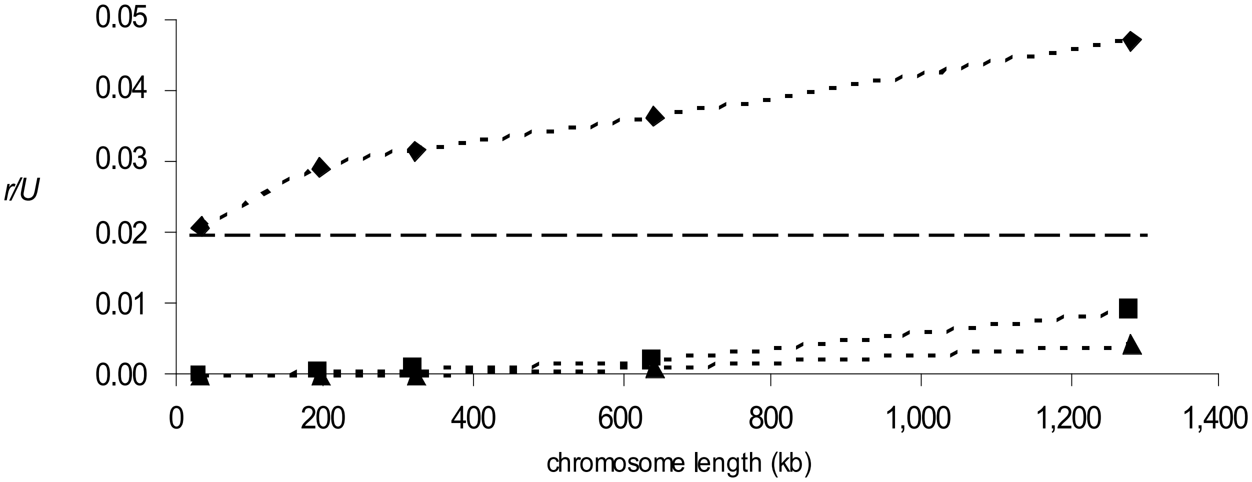


Figure 3-4

The effect of increasing the number of sites subject to BGS on the rate of fixation of “major” mutations. **Symbols:** Diamonds: *U* = 3.78 and *s* = 0.63, which correspond to the scaled values of *U* = 0.009 and *s* = 0.0015 as estimated for the *D. miranda* neo-Y. Selection at the BGS sites is drawn from a log-normal distribution with *Ns*(harmonic mean) = 10. The dashed line indicates the observed *r/U* when there is no BGS for *U* = 3.78, *s* = 0.63.

Squares: *U* = 2.50, *s* = 0.63; triangles: *U* = 1.89, *s* = 0.63.

3.5 Discussion

3.5.1 Plausibility of the model

The importance of Muller's ratchet in driving the degeneration of Y chromosomes (CHARLESWORTH and CHARLESWORTH 1978; GORDO and CHARLESWORTH 2000a, b and 2001) has been questioned, mainly because the time-scales involved were inferred to be too large to be biologically significant (CHARLESWORTH 1996). In addition, ENGELSTÄDTER (2008) showed that, under some circumstances, the presence of deleterious mutations on X-linked homologues of the Y chromosome can greatly slow down the ratchet, compared with what is found in haploid simulations of the type used here.

We first consider this technical problem, and then discuss the question of the rate of the ratchet in relation to Y chromosome evolution. The selection coefficients used in our models of major mutations is the harmonic mean selection coefficient against major mutations on the evolving neo-Y chromosome, which we obtained from LANGLEY *et al.* (1981), who used data on the frequencies of null alleles at autosomal loci in *D. melanogaster* populations together with the rate of mutation to null alleles with the the standard formula for mutation-selection equilibrium (HALDANE 1927). Since frequencies of null alleles are very low, this estimate can be equated to the harmonic mean of $hs' + q^*s'$, where h is the dominance coefficient, s' is the homozygous selection coefficient, and q^* is the equilibrium frequency of a null allele at a locus; the latter term takes into account the contribution of the occasional homozygote to the net fitness of a null allele. Under the assumptions of our model, this quantity should be the same as the harmonic mean selection coefficient experienced by a major mutation on the neo-Y chromosome, taking into account the presence of rare allelic mutations on the neo-X chromosome, so that our haploid model should accurately represent the early evolution of the *D. miranda* neo-Y chromosome.

Our simulation results show that a high rate of accumulation of strongly deleterious mutations on an evolving Y chromosome in a Drosophila population can be achieved with biologically reasonable parameters, due to Hill-Robertson interference effects among sites subject to purifying selection. Under free recombination, the expected rate of fixation of strongly deleterious loss-of-function mutations (equation 3) is virtually zero when $s = 0.63$, $U = 3.78$ and $N = 1{,}000$. When rescaled to a haploid population size of 210,000, these correspond to a harmonic mean selection coefficient against major mutations of 0.0015, and a mutation rate to major mutations of 0.009, which we have argued are plausible values for the neo-Y chromosome of *D. miranda*. However, with no recombination, such mutations can become fixed in our simulations, at a rate that is similar to the rate observed for the neo-Y, especially when we take the effect of weak selection against amino acid mutations in the background ("BGS sites") into account.

Given the exponential dependence of the rate of the ratchet, r, on U, a reduction in U has a large effect on r (equations 1 and 2). As about half of the genes originally present on the *D. miranda* neo-Y have lost their function since the origin of the chromosome (BACHTROG *et al.* 2008), the maximum reduction in U we can assume is 50%. Figure 3.4 shows that, when U is reduced to 2/3 or 1/2 of its original value, r/U is indeed reduced, probably below the observed value of 9%.

It is important to note in this context that the relative effect of selection at BGS sites on the speed of the ratchet is actually larger when U is smaller: for $U = 2.5$, r/U increases by a factor of about 160 when the number of BGS sites increases from 32 kb to 1.28 Mb, as opposed to a factor of only about 2-4 for $U = 3.78$. When $U = 1.89$, the ratchet has stalled for short chromosomes ($L = 32$ kb), but is going at a relative rate of 0.4% for $L = 1.28$Mb.

As previously suggested (BACHTROG 2008b; ENGELSTÄDTER 2008), the speed of the ratchet is likely to vary at different stages of Y degeneration: when the overall occurrence of major mutations is still high, interference among very strongly selected mutations alone leads to their fast accumulation, and the process is accelerated (about 2-4 fold) by the presence of BGS sites. With the erosion of genes from the neo-Y, U decreases and the ratchet slows down, but the effect of mutations at nonsynonymous sites starts to increase, until U becomes so low that the BGS effect cannot greatly increase the ratchet any longer (eventually the BGS effect will disappear as well). This process might lead to a stable situation, i.e. once the Y chromosome contains few enough genes, selection will be able to purge the occasional loss-of-function mutations that hit it.

Indeed, most old non-recombining chromosomes, such as the Y chromosome in humans or *D. melanogaster*, or the chicken W are very small (CARVALHO *et al.* 2009; SKALETSKY *et al.* 2003), and Muller's ratchet may no longer be driving the fixation of knock-out mutations. However, the process *leading* to this stage from a large non-recombining region of the genome may well have been driven by a ratchet. The effects of the ratchet could also be a factor limiting the size of non-recombining regions of the genome, and, consistent with this, major gene content expansion of a non-recombining region has not yet been reported, although small numbers of genes have been added to the highly degenerated Y chromosomes of Drosophila and mammals (CARVALHO *et al.* 2009; SAXENA *et al.* 1996). (The successive expansion of the non-recombining regions of evolving Y chromosomes, resulting in "evolutionary strata" (LAHN and PAGE 1999), is a quite different process from this, since it involves a succession of events that create newly non-recombining genomic regions from previously recombining ones.)

3.5.2 The effects of dosage compensation

In an evolving sex chromosome system, gene loss from the Y chromosome is expected to lead to the evolution of dosage compensation (CHARLESWORTH 1978) since there is a selective advantage to increase the expression of functional alleles on the neo-X relative to their inactivated counterparts on the neo-Y. Even though the exact mechanisms are still unknown, there is evidence for partial dosage compensation of the *D. miranda* neo-X/neo-Y (BACHTROG 2006; BONE and KURODA 1996; MARIN *et al.* 1996; STEINEMANN and STEINEMANN 1999)

There is no clear relationship between the rate of amino acid evolution on the neo-Y and the relative expression levels of neo-Y versus neo-X genes (BACHTROG 2006), which generally but not always have lower expression than their X-linked counterparts. It is not in fact clear that this reduction in neo-Y gene expression reflects dosage compensation; the fact that a minority of neo-Y genes are more highly expressed than their counterparts on the neo-X suggests that mutations in regulatory sequences that disturb expression in either direction may well be accumulating (BACHTROG 2006). Such mutations could increase interference effects relative to the scenario where loss-of-function mutations are only occurring in coding regions, and hence speed up the ratchet. Similarly, genes that are recruited onto the Y chromosome (and hence not present on the X) will be under strong purifying selection and contribute to *U*. The overall impact of these factors seems, however, unlikely to change the parameter space to such an extent that the ratchet would come to a halt during the early stages of neo-Y evolution, although the ratchet may well not account for the full degeneration of Y chromosomes (BACHTROG 2008b).

3.5.3 The effect of BGS sites on the rate of the ratchet

When the difference in mean *s* between the two types of selected sites (BGS sites and major sites) is very large, the model based on the reduction in N_e deduced from neutral sites subject to BGS performs poorly in predicting the rate of the ratchet, i.e. the ratchet clicks a lot more slowly than expected. In other words, the pool of individuals from which the population is ultimately derived, the "least-loaded class", is larger than that predicted from the effective number of individuals that determines levels of nucleotide diversity in the BGS simulations (chapter 2) (Figures 3-3 and 3-4). This makes intuitive sense, because the important factor determining transmission to the next generation is fitness relative to the population average. The relative fitness reduction due to "major" mutations is about 30 times larger than that mutations at BGS sites for the parameters used in Figures 3-4; hence, a few mutations at BGS sites in a chromosome will not make much of a difference for an individual that is otherwise free of "major" mutations– its overall fitness will still be high compared to the rest of the population. However, when the *s* values at the two types of sites are similar (as in Figure 3-2), mutations at BGS sites can reduce the chance of reproduction considerably. In addition, the movement of a ratchet is associated with a substantial reduction in N_e at neutral or weakly selected sites (GORDO *et al.* 2002). This means that the N_es values at the BGS sites in our simulations will be greatly reduced relative to their values in the absence of the major mutations, thereby undermining their ability to cause Hill-Robertson interference. When *U* at the "major" sites is reduced, the relative effect of BGS mutations becomes larger, increasing the impact of BGS on the ratchet (Figure 3-4).

The idea that that the extent of interference with other sites on the speed of the ratchet is affected by the relative magnitudes of the selection coefficients at the sites concerned has previously been discussed in a somewhat different context (GORDO and CHARLESWORTH 2001; SÖDERBERG and BERG 2007).

3.5.4 Can selective sweeps alone explain the neo-Y data?

In principle, selective sweeps can drag to fixation deleterious mutations, as long as the fitness benefit due to the advantageous mutation outweighs the cost of carrying deleterious mutations in the genomic background (CHARLESWORTH 1994; HADANY and FELDMAN 2005; JOHNSON and BARTON 2002; RICE 1987). Recent studies of DNA sequence evolution in *D. melanogaster* and *D. simulans* have suggested that around 50% of fixed differences between species in protein sequences and some types of non-coding sequences such as UTRs are the result of positive selection (e.g. (ANDOLFATTO 2007). If this applies to *D. miranda* and its relatives, as is suggested by recent data ((BACHTROG 2008a); HADDRILL et al., unpublished data), then there has been ample opportunity for numerous selective sweeps on the evolving neo-Y chromosome lineage, given the size of this chromosome (around 3,000 genes) and its time of origin (1.75 million years ago, corresponding to $K_S = 0.01$). For example, with $K_A / K_S = 0.08$ and assuming 1000 nonsynonymous nucleotide sites per gene, as is typical for Drosophila proteins (CLARK *et al.* 2007), we would expect approximately 0.5 x 0.08 x 0.01 x 1000 x 3000 = 1200 sweeps to have occurred if the neo-Y protein sequences were evolving at the standard rate. With 10 generations a year, this would correspond to 1 sweep every 14,600 generations on the neo-Y lineage, a relatively modest rate.

Given this low rate, it is reasonable to assume that each sweep is followed by a period of recovery, after which mutation-selection balance for major mutations would be approximately re-established (ignoring the ratchet). The estimates for mutation and selection against loss-of function mutations on the neo-Y (see above) suggest that each ancestral neo-Y would have carried, on average, $U/s = 6$ major deleterious mutations. Hence, an advantageous mutation will, on average, arise on a genomic background that carries six such mutations. Assuming that there were originally about 3,000 genes on the neo-Y, about half of which now carry "major" mutations, this suggests that the number of sweeps necessary to explain the data is about 1,500/6 = 250, much smaller than the above estimate.

This is, however, likely to be a conservative estimate for several reasons. First, even though an advantageous mutation will, on average hit a chromosome carrying six major deleterious mutations, the chance of fixation of the beneficial mutation will be higher when it happens to hit a chromosome with fewer mutations, decreasing the average number of "major" mutations fixed with each sweep. Second, with a decline in *U* over time, the number of "major" mutations segregating in the population will also decline, and with it the number of deleterious mutations fixed with each sweep. (On the other hand, the chance of fixation of the beneficial mutation will increase when there are fewer deleterious mutations in the background, so these effects might weigh each other out).

Third, the fixation probability of a beneficial mutation can be greatly reduced by selection acting against linked deleterious mutations, unless the selection coefficient of the beneficial mutation is larger than *U* (Figure 2 and equation C9 in JOHNSON and BARTON (2002)). In the present case, this result suggests that only beneficial mutations with selective advantages of the order of 1% would have a reasonable chance of fixation. The two methods that have been used to estimate these selection coefficients in Drosophila yield very different values: 1% versus 10^{-5} (SELLA *et al.* 2009). The latter value would clearly not be compatible with the fixation of beneficial mutations in the presence of the major mutations we have been considering, in contrast to the former value. However, the method giving an estimate of 1% for the selection coefficient of beneficial mutations gives an estimate of the rate of sweeps that is 10% of the value we are using here (SELLA *et al.* 2009), corresponding to only 120 sweeps since the origin of the neo-Y chromosome. It thus seems unlikely that we can ascribe the fixation of all major mutations to the effects of sweeps, but we cannot exclude the possibility that selective sweeps have contributed to some of the observed fixations, although patterns of neutral diversity on the neo-Y are consistent with the action of purifying selection alone (chapter 2). It is also worth mentioning that a recent analysis of the non-crossing over dot chromosome of *D. americana*, which has only 80 genes on it, found no evidence for adaptive fixations of amino acid mutations, in contrast to the large fraction of fixations on other chromosomes that appeared to have been driven by positive selection (BETANCOURT *et al.* 2009). This strongly suggests that the reduction in effective population size associated with reduced recombination greatly reduces the efficacy of positive selection, which would undermine the ability of selective sweeps to contribute to the degeneration of Y chromosomes.

4 *Slcyt*, a newly identified sex-linked gene, has recently moved onto the X chromosome in *Silene latifolia* (Caryophyllaceae)

This work has been published in the following paper:

KAISER, V. B., R. BERGERO and D. CHARLESWORTH, 2009 *Slcyt*, a newly identified sex-linked gene, has recently moved onto the X chromosome in *Silene latifolia* (Caryophyllaceae). Molecular Biology and Evolution 26: 2343-2351

Contributing Authors:

- I carried out the laboratory work, the analysis and writing of the manuscript
- R. Bergero helped with the laboratory work, the analysis and writing of the manuscript
- D. Charlesworth helped with the analysis and writing of the manuscript
- Plant material was made available from collections of R. Bergero and D. Charlesworth
- R. Bergero handled the greenhouse and DNA extractions and constructed the cDNA libraries
- Roberta Bergero and Helen Borthwick mapped genes in *S. vulgaris*, except for the homologues of *SlX9Y9* and *SlCyp*
- Helen Borthwick carried out the DNA extractions and assisted in the laboratory
- Sequencing and capillary electrophoresis were performed by the Edinburgh University Sequencing Service

4.1 Abstract

The sex chromosomes of the plant species *Silene latifolia* (white campion) are very young (only 5-10 My old), and all eleven X-linked genes so far described have Y-linked homologues. Theory predicts that X chromosomes should accumulate a non-random set of genes. However, little is known about the importance of gene movements between the X and the autosomes in plants, or in any very young sex chromosome system. Here, we isolate from cDNA a new gene, *Slcyt*, on the *S. latifolia* X, which encodes a cytochrome B protein. We genetically mapped *SlCyt* and found that it is located ~1cM from the pseudoautosomal region. Genes in this region of the X chromosome have low divergence values from their homologous Y-linked genes, indicating that the X only recently stopped recombining with the Y. Genetic mapping in *S. vulgaris* suggests that *Slcyt* originally belonged to a different linkage group from that of the other *S. latifolia* X-linked genes. *S. latifolia* has no Y-linked homologue of *Slcyt*, and also no autosomal paralogues seem to exist. *Slcyt* moved from an autosome to the X very recently, as the *Cyt* gene is also X-linked in *S. dioica*, the sister species to *S. latifolia*, but is probably autosomal in *S. diclinis,* implying that a translocation to the X probably occurred after the split between *S. diclinis* and *S. latifolia/S.dioica*. Diversity at *Slcyt* is extremely low (π_{syn} = 0.16%), and we find an excess of high-frequency derived variants and a negative Tajima's *D*, suggesting that the translocation was driven by selection.

4.2 Introduction

In the evolution of animal sex chromosomes, gene movements occur both from and to the X chromosome. There are grounds to believe that this is not random, but that selection acts to enrich the X (or Z in ZW systems) for genes with sex-specific functions, especially sexually antagonistic genes, i.e. genes that increase fitness in one sex but are deleterious in the other sex (RICE 1984). In mammals and Drosophila, the X chromosome has a non-random gene content, having an overrepresentation of male-specific genes in mammals (KHIL *et al.* 2004; LERCHER *et al.* 2003; MUELLER *et al.* 2008; WANG *et al.* 2001), and an under-representation in

Drosophila (OLIVER and PARISI 2004; PARISI *et al.* 2003; VICOSO and CHARLESWORTH 2006); in the chicken, female-specific genes are underrepresented on the Z chromosome (KAISER and ELLEGREN 2006). Non-random distributions of sex-biased genes could evolve through the evolution of biased expression of initially non-biased genes that were already located on the X. Alternatively, genes could be recruited onto the X from the autosomes. For instance, the *Drosophila melanogaster* Y carries mainly male function genes, none of which has an X-linked homologue (BROSSEAU 1960; CARVALHO 2002). It could be advantageous for a female-beneficial antagonistic gene to move to the X if the amount of gene product is directly related to its copy number, since such a movement would lead to lower average expression in males. However, this will depend on several factors, including details of the dosage compensation system. If dosage compensation occurs, and it acts on the whole chromosome or large X regions, such a change could be either selectively neutral or deleterious. If, however, expression in the two sexes is equalized by up-regulating the X chromosome in males, as in Drosophila (GUPTA *et al.* 2006), genes moved onto the X will not gain increased relative female expression. Finally, if dosage compensation occurs in females, for example by inactivating one X, and increasing expression from the other X, as in mammals (NGUYEN and DISTECHE 2006), translocating female-beneficial genes to compensated regions of the X might be disfavoured in females, because expression relative to autosomal genes will often be reduced.

Both mammals and Drosophila have genetically degenerated Y chromosomes, and gene movements to and from their X chromosomes have been documented (BETRAN *et al.* 2002; EMERSON *et al.* 2004). These movements occurred over long evolutionary times, during most of which dosage compensation operated for at least some X-linked genes. The mammalian sex chromosomes evolved about 170 MYA, after the marsupial and Eutherian mammal lineages split from the ancestor of the platypus (VEYRUNES *et al.* 2008), consistent with the very high X-Y sequence divergence of a few genes (LAHN and PAGE 1999; ROSS *et al.* 2005). In Drosophila, the XY system is at least 39MY old, and in *D. melanogaster* the X is Muller's chromosome element A, the ancestral Drosophila X chromosome

(ASHBURNER *et al.* 2005; CARVALHO 2002). The bird sex chromosomes also stopped recombining very long ago (reviewed in NAM and ELLEGREN (2008)).
Studying the evolution of the X gene content in younger sex chromosome systems should shed light on how this content evolves, and plants are of interest for such studies. The highest silent site divergence values between homologous X- and Y-linked genes of *Silene latifolia* (Caryophyllaceae) are just over 20%, suggesting that the sex chromosomes evolved only about 5-10 MYA (BERGERO *et al.* 2007; FILATOV 2005; NICOLAS *et al.* 2005). Recombination in the regions nearest to the pseudoautosomal region (PAR) ceased much more recently than in other regions, and several genes near the *S. latifolia* PAR have X-Y silent site divergence below 5% (BERGERO *et al.* 2007; NICOLAS *et al.* 2005). Similarly, in mammals, genes on the X and Y fall into four or five "strata" of diminishing evolutionary ages as their locations get closer to the PAR, suggesting step-wise or gradual recombination suppression, though on a time-scale much longer than in *S. latifolia* (LAHN and PAGE 1999; NAM and ELLEGREN 2008). These findings are consistent with the hypothesis that, in both these X chromosomes, sexually antagonistic genes may have accumulated over time, leading to selective pressure for suppressed recombination in regions where recombination formerly occurred, to maintain associations between such genes and the sex-determining genes.

However, this is indirect evidence, as is all other currently available evidence for sexually antagonistic genes on evolving sex chromosomes (ELLEGREN and PARSCH 2007; MANK and ELLEGREN 2009; MANK *et al.* 2008). One reason for studying plant sex chromosomes is the hope that such genes may be discovered. Dioecious plants have few secondary sexual differences, but flowers and inflorescences of males and females often differ (LLOYD and WEBB 1977). In *S. latifolia*, the two sexes differ in their optimum numbers and sizes of flowers, leading to a genetic conflict, and females with fewer, larger flowers produce sons that have fewer, larger flowers than the average male (DELPH *et al.* 2004; DELPH *et al.* 2002; PRASAD and BEDHOMME 2006). In other dioecious plants, traits not associated with flower morphology also differ between the sexes, e.g. leaf resin content, leaf and stem morphology, senescence patterns and plant-herbivore interactions (CORNELISSEN and STILING 2005; KRISCHIK and DENNO 1990; MERZOUKI *et al.*

1996), suggesting that genes expressed differentially in males and females may be common in dioecious plants. Currently, however, sex-specific expression in plants is known only for genes expressed exclusively in reproductive tissue (SATHER *et al.* 2005; YU *et al.* 2008).

The gene content of plant sex chromosomes is also currently largely unknown, except for the Y chromosomes of the moss *Marchantia polymorpha* (ISHIZAKI *et al.* 2002) and *S. latifolia*. In *S. latifolia*, eleven Y-linked genes have now been described, all apparently functional, except for the incomplete *MROS3*-Y gene (GUTTMAN and CHARLESWORTH 1998), and possibly *SlssY* (FILATOV 2008). Here, we describe the first case of a movement to the *S. latifolia* X. The gene, *Slcyt*, was recently translocated from an autosome in *S. latifolia*, and inserted close to the pseudoautosomal region of the X chromosome. This rearrangement could have led to suppressed recombination in the region, and, as we discuss below, the movement might have been driven by sexual antagonism. The *Slcyt* gene seems to have been affected by a recent selective sweep. The only other gene movement known in a plant sex chromosome system is the duplicative transposition of a gene to the Y, and the Y copy has increased expression in stamens, compared with the autosomal one, suggesting that this was probably also an advantageous gene movement (MATSUNAGA *et al.* 2003). Recently, it has also been discovered by cytogenetic studies that *S. diclinis* has an X-autosome translocation not present in *S. latifolia* (HOWELL *et al.* 2009). No genes in the added region have yet been studied. Our present study suggests that the X-autosome translocation involving the *Slcyt* gene is a separate event from that in *S. diclinis*. It has so far been generally accepted that no additions to this sex chromosome pair had occurred during its evolution, but that it evolved from a single ancestral autosome which can be identified by its gene content (FILATOV 2005), but these new results show that at least part of the X near the PAR has recently been added to one or both of the XY chromosome pair.

4.3 Methods

4.3.1 *Silene* DNA samples

The study used *S. latifolia*, *S. dioica*, *S. diclinis* and *S. vulgaris* plants from natural populations, which are described below. *Silene dioica* and *S. diclinis* have the same XY system as *S. latifolia,* i.e. genes in the "older" regions of the *S. latifolia* Y chromosome stopped recombining with the X before this group of dioecious species split; the sequences of these genes form distinct X and Y clusters, rather than clustering by their species of origin (NICOLAS *et al.* 2005). *S. dioica* is the sister species to *S. latifolia*, and these species often hybridize in nature (BAKER 1948; FILATOV *et al.* 2001; LAPORTE *et al.* 2005). *S. vulgaris*, the outgroup species used in this study, is gynodioecious and has no sex chromosomes (DESFEUX *et al.* 1996). Individuals were grown in the greenhouse at the University of Edinburgh and DNA was extracted from fresh leaves using the Fast DNA kit (Q-biogene), following the manufacturer's instructions which can be found at www.qbiogene.com.

Sex-linked and autosomal genes were identified in the mapping families F2005-4 and H2005-1, which are F2 families descended from crosses between geographically distant populations (BERGERO *et al.* 2007). For putatively X-linked genes, we genotyped the two F1 individuals (parents of the F2) of H2005-1, and scored 92 F2 offspring for variants found in the maternal and/or paternal plant. For each gene, the inheritance patterns of the two variants within the family were compared to the pattern obtained for previously published X-linked genes.

For mapping genes in *S. vulgaris*, two families (named SV1 and SV2) were used, with 51 and 64 offspring respectively. These families were generated by crossing a female plant (E2000 5/9, from Dijon, France) with two unrelated hermaphroditic plants (H2000-4 and 99K-10-4, from Flamanville, France, and Sussex, England, respectively). A linkage map of the markers was inferred using the software JoinMap (STAM 1993).

4.3.2 Identifying sex-linked genes

A set of *S. latifolia* gene sequences were isolated from a cDNA library constructed by a simplified version of the template-switching (TS) procedure of MATZ *et al.*

(1999). First strand cDNA was synthesized from total RNA extracted from male leaf primordia, using 50 mM of oligo(dT)$_{21}$ (5′-GATCGATTTTTTTTTTTTTTTTTTTTTVN-3′), 30 mM $MgCl_2$ and 200 U reverse transcriptase Superscript II (Invitrogen, Paisley, UK), following the manufacturer's recommendation specified at www.invitrogen.com. The TS adapter (5′-GGTTTTGGTAGTTCTGTGTGTTGGG-3′) was ligated to the 5′ ends of cDNAs in a 50 µl reaction containing 5 units of Klenow-fragment 3′ → 5′ exo^{-} (New England Bioloabs, Ipswich, MA), 1x buffer 2 (New England Biolabs, Ipswich, MA), 1 mM dNTPs, and 50 picomoles of TS adapter. The reaction was carried out at 16 °C, overnight. After incorporation of the TS adapter, the cDNA was purified on a QIAGEN spin column, polymerase chain reaction (PCR) amplified using the primer pair for the TS-adapter and an oligo(dT)$_{21}$ primer, and finally cloned in a T-tailing pBS vector (Stratagene, La Jolla, CA). We refer to this as the TS library.

Candidate sex-linked genes in *S. latifolia* were identified using a combination of segregation analysis of intron size variants (ISVS), using a universal primer (5′-GGTTGGAGCTAGTGTTGTG-3′) labelled with 6-FAM or VIC (Applied Biosystems, Foster City, CA), and direct sequencing, as described by BERGERO *et al.* (2007). Briefly, we first identified putative intron positions by comparing the *S. latifolia* cDNA sequences with the translated *A. thaliana* and *O. sativa* (rice) genome sequences, using BLASTX at www.ncbi.nlm.nih.gov. PCR primers were then designed from the *S. latifolia* cDNA sequences flanking putative introns, using the Oligonucleotide Properties Calculator available at http://www.basic.northwestern.edu/biotools/oligocalc.html. The PCR conditions using the labelled universal forward primer were generally as follows: 10 cycles of 94°C for 30sec, 56°C for 30sec, 72°C for 1min, followed by 25 cycles of 94°C for 30sec, 50°C for 30sec, 72°C for 1min; final extension at 72°C for 30min. PCR conditions without the universal labelled forward primer: 10 cycles of 94°C for 30sec, 56°C for 30sec, 72°C for 1min, followed by 25 cycles of 94°C for 30sec, 52°C for 30sec, 72°C for 1min. To detect size differences that could be used as genetic markers, the PCR products were run on 1.5% agarose gels and inspected visually.

For sequences that did not yield suitable size variants, capillary electrophoresis was performed on an ABI 3730 capillary sequencer (Applied Biosystems, Foster City, CA). The labelled forward primer allowed length variants among the PCR products to be scored using the Genemapper software package 3.7 (Applied Biosystems, Foster City, CA). PCR amplicons that showed no length variants were directly sequenced, and sequences were examined in Sequencher 4.5 to detect variants suitable for segregation analysis after digestion with restriction enzymes. Finally, if no suitable restriction sites were found, genotyping for segregation analysis was performed by direct sequencing to detect polymorphic variants. A total of sixteen genes were tested for sex-linkage.

4.3.3 Obtaining the complete *Slcyt* sequence

Since the cDNA sequence that allowed us to determine that *Slcyt* is sex-linked (see Results) did not contain the whole *Slcyt* coding sequence, nested PCR was used to obtain the 5´ end of the gene. The TS cDNA library was used as template in a first round of PCR using a primer for the TS-adapter and the reverse gene-specific primer *Slcyt*_b_R (Table 4.1). One microlitre of the first-round PCR was used in a second round PCR with primer for the TS-adapter and the gene-specific primer *Slcyt*_R, which binds internally to the PCR product from *Slcyt*_b_R (Table 4.1). The PCR products were cloned, sequenced in an ABI 3730 capillary sequencing machine (Applied Biosystems) and visualized using Sequencher 4.5 software.

Intron 1 proved difficult to amplify, but was amplified with Phusion enzyme (Finnzymes, Espoo, Finland) and a PIKO 24 thermal cycler (Finnzymes, Espoo Finland), following the manufacturer's instructions (PCR conditions: 40sec at 98°C, followed by ten cycles of 5sec at 98°C, 5sec at 62°C, 5min at 72°C; 25 cycles of 5sec at 98°C, 5sec at 55°C, 5min at 72°C; final extension at 72 °C for 5min).

4.3.4 Analyses of DNA sequence diversity

To study DNA sequence diversity of *Slcyt* within *S. latifolia*, parts of the gene (starting within exon 2, and including the whole of intron 2 and exon3, and parts of the 3´ UTR, see Figure 4.1) were amplified in the same 48 European male plants, from 24 different European populations, covering the entire native range of the

Table 4.1: Primers used in order to amplify *Slcyt* and *SlX9*

Name	**Sequence**	**Notes**
*SlX9*_F (exon2) *SlX9*_R (exon3)	CTTGTGGAACTTCTGGTGGAAG GTCCAATCACATTCAAGTCTCTCC	establishing sex linkage of *SlX9* in *S. latifolia*; mapping in *S. latifolia* and *S. vulgaris*
*Slcyt*_F2(exon2) Slcyt_b_R (exon 3)	GAGATGATGTCTTCCTTGATGC TAGAGGAAGCCTACTATGACAGC	establishing sex linkage of *Slcyt* in *S. latifolia*
*Slcyt*_ex2_cons_F *Slcyt*_3_R (3'UTR)	ACCCCGGTGGAGATGATG CAACTTCTTGTCAAAATTGATCG	Studying diversity of *Slcyt*
Slcyt-I1-F-univ.-1 *Slcyt*-I1-R-2	GTCTGGAGCTAGTGTTGTGTTCATCTGCTCTTCATATTCTTCG AATTGGGAATAGGGATGCATTTGC	Mapping *Slcyt* in *S. latifolia* ; amplifying SlX_STR1in *S. latifolia*, *S.dioica* and *S. diclinis* (forw. primer only).
Cyt-dicl-I1-R	TTCTACTTGAGACCACAAATTCTC	amplifying SlX_STR1in *S. diclinis*

Table 4.1, continued

Cyt-exon1-F-1	CGAAACTGGTTAGTATGCAAGAAG	Mapping *Cyt* in *S. vulgaris*
Cyt-exon2-R	TCAAGGAAGACATCATCTCCAC	
Slcyt-E3-UTR-R	GAGTTTCCTATTTGCGCAAGTAGAG	Other *Slcyt* primers
*Slcyt*_R2(exon3)	TAGAGGAAGCCTACTATGACAGC	
*Slcyt*_ex3_cons_R	GCATCGTCAAATTCTTCTTTTGCATC	
*Slcyt*_F (exon 3)	AGATGCAAAAGAAGAATTTGACG	
Slcyt-E3-beg2-R-univ	GTCTGGAGCTAGTGTTGTGTGTTGGAAGGA ATTTGAGG	
Slcyt-E3-beg-R-univ	GTCTGGAGCTAGTGTTGTGCCAGCATCGTCAAATTCTTC	
*Slcyt*_ex1_cons_F:	TCTAAGGATGATTGTTGGGTTGTC	
Slcyt-exon1-F-2	TGTTGGGTTGTCATTCATGG	
*Slcyt*_R(exon3)	ACAGCGATACAAACAACAGCAC	

species (the samples are listed in Table 4.2) that were used to confirm the absence of a Y-linked copy (see Results), using the primers *Slcyt*_ex2_cons_F and *Slcyt*_3_R (3'UTR) (Table 4.1). The PCR products were directly sequenced and edited in Sequencher. Sequence diversity was analyzed using DnaSP 4.0 software (ROZAS and ROZAS 1999), which was also used for several tests of neutrality, including Tajima's D, Fu and Li's D* and F*, Fay and Wu's H and Fu's F statistics (FAY and WU 2000; FU 1997; FU and LI 1993; TAJIMA 1989). The significance levels for these tests were calculated using coalescent simulations implemented in DnaSp, assuming no recombination (FILATOV 2008), which is a conservative approach (TAJIMA 1989; WALL 1999). The HKA test (HUDSON *et al.* 1987) was used to compare *Slcyt* diversity levels with those of *X7*, *Cyp-X, X4* and *SlX9*, another new X-linked gene, which will be described in Chapter 5.

In the course of studying sequence the diversity of *Slcyt*, we identified a polymorphic $(TTA)_n$ microsatellite in intron 1 (referred to below as SlX_STR1). This was scored using the primers *Slcyt*-I1-F-univ.-1 and *Slcyt*-I1-R-2 (Table 4.1), and the sizes of the amplicons were determined using GeneMapper 3.7. Ten out of eleven females from different natural populations were heterozygous for SlX_STR1 (Table 4.3). To further test for complete sex-linkage (and exclude a pseudo-autosomal location for *Slcyt*), SlX_STR1was amplified in the 48 males listed in Table 4.2. Finding heterozygous males would indicate the presence of a Y-linked copy, or a pseudoautosomal location.

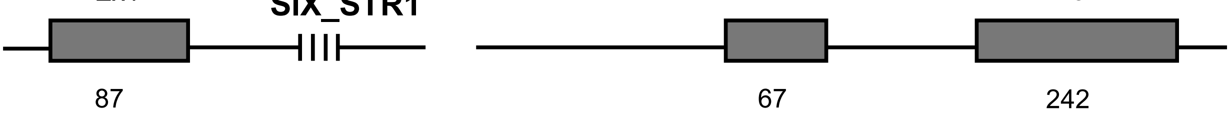


Figure 4.1: Gene structure of *Slcyt*. Sizes in bp of exons, and the position of the microsatellite SlX_STR1are given. There is a gap in the sequence of about 70bp in intron 1, which has a total length of about 730bp.

Table 4.2: Origins of male individuals of the *Slcyt* diversity study, and the length of the region that includes the microsatellite SlX_STR1 in intron 1 of *Slcyt*. The amplification of SlX_STR1was not successful for individuals no. 8 and 37. Non-integer numbers of basepairs must be due to inaccuracies of length scoring by GeneMapper 3.7

ID	Length of amplicon (bp)	Country	Population
1	161	Germany	Bissendorf
2	176		
3	191	Sweden	Oland, Grasgard
4	191		
5	164	France	Relais des Chenes
6	176		
7	148.5	Greece	Ioanninon
8	-		
9	158	Poland	Krakow, near Conference Centre
10	155		
11	167	Denmark	Aarhus Botanic Garden, origin unknown
12	165		
13	213	Austria	Vienna, Heiligenstadt
14	213		
15	185	Norway	Lardalsoyri
16	164		
17	145	Greece	Evrou
18	173		

Table 4.2, continued

19	155	Italy	near Tarquinia, Lazio
20	155		
21	161	Estonia	Vinupea
22	170		
23	145	Austria	Dietmans
24	145		
25	191	France	Canche
26	158		
27	158	Portugal	Serre de Nogere
28	148.5		
29	148.5	Netherlands	River Kraal
30	138		
31	138	France	Vitry en Artois
32	129		
33	129	Italy	Piedmont, Ceres
34	142		
35	155	UK	Dalkeith
36	185		
37	-	UK	Somerset, Burnham-on-Sea
38	185		
39	119.5	Russia	Krasnoyarsk, Siberia

Table 4.2, continued

40	119.5	Russia	Krasnoyarsk, Siberia
41	167	Portugal	Segier-Chavez
42	148.5		
43	155	Spain	Madrid, San Lorenzo de El Escorial
44	148.5		
45	161	Ukraine	Ukraine
46	164		Ukraine/Belarus border
47	185.5	Germany	near Glaubitz
48	185.5		

Table 4.3: Lengths of the region that includes the microsatellite SlX_STR1 in intron 1 of *Slcyt* in female *S. latifolia* individuals from various natural populations.

Individual	Location	Length of amplicon (bp)
K2005-2B	France	171, 176
K2005-1B2	Germany	161, 176
G2005-5/3	Greece	148, 161
G2005-4	Greece	135, 145
J2006-1/3	Denmark	159, 177
E2004-17/3	Netherlands	149, 177
K2005-3A/3	Poland	155, 171
K2005-4/3	Austria	174, 214
K2005-9/3	Estonia	162, 171
K2005-2A/3	France	165, 222
K2005-7/3	Norway	165

4.4 Results

4.4.1 Identification of two new X-linked genes in *S. latifolia*

A total of 16 *S. latifolia* cDNA sequences were screened for sex-linkage using the ISVS method, direct sequencing and/or restriction digestion (see Methods). The PCR primers used for mapping are listed in Table 4.1. Two new genes were found to be sex-linked, giving a total of thirteen genes so far identified on the X and/or Y chromosome in this species. They were provisionally named *Slcyt* and *SlX9/SlY9*.

Sex-linkage of the first gene, *Slcyt*, was established by direct sequencing and segregation analysis of an SNP variant (G/C) in intron 2. The male parent of the F2005-4 mapping family previously studied (BERGERO *et al.* 2007) carried a G while the female parent was homozygous for C. Seven male offspring were scored, and all inherited the mother's variant, whereas all four female offspring scored were heterozygous for both variants, strongly suggesting that the gene is X-linked. Further evidence for X-linkage is given below.

The other new sex-linked gene, *SlX9,* was mapped using the mapping family H2005-1 (BERGERO *et al.* 2007), by scoring a size variant in intron 2 directly on an agarose gel. Variants in the *SlX9/SlY9* gene found exclusively in the father and male offspring (Y-linked sequences) were identified using segregation analysis of a marker that was heterozygous in both the maternal and paternal plants. This gene will be described in more detail in chapter 5.

4.4.2 Genetic mapping of *Slcyt* and *SlX9*

Slcyt and *SlX9* were genetically mapped using segregation analysis among the 92 F2 offspring of the mapping family H2005-1. The maternal plant was heterozygous for both genes. For *Slcyt*, the F2 offspring were scored for SlX_STR1. Again, there were no heterozygous males among the offspring, and no recombinants were found between *Slcyt* and *SlX6b* among the 92 offspring scored, supporting the conclusion that *Slcyt* is X-linked, and is located in the part of the X that does not recombine with the Y in males. *SlX6b* maps at a distance of ~1cM from the pseudoautosomal marker, OPA (BERGERO *et al.* 2007). One of the mother's two SlX_STR1 alleles appeared in only two out of 43 males and four out of 49 females.

This aberrant ratio ($\chi^2 = 93.56$, $p < 0.001$, 1d.f.) is similar to results for *SlX6b* (R. Bergero, unpublished data), and suggests either a bias in transmission via female gametophytes or higher mortality of offspring carrying this maternal X chromosome. No such bias was detected in five other mapping families tested, but these families confirmed X-linkage of *Slcyt*. No Y-linked copy was detected (see below).

The same intron size variant that was used to establish sex-linkage was also used to map *SlX9* in relation to the other known X-linked genes. *SlX9* maps to the same position as *SLCyp*-X, a previously described sex-linked gene located about 14 cM from the pseudoautosomal marker (BERGERO *et al.* 2007).

4.4.3 Mapping orthologues of X-linked genes in *Silene vulgaris*

Because no Y-linked copy was detected, we tested linkage of the *Slcyt* orthologue in *S. vulgaris* (which we denote by *Svcyt*), to test whether the gene has changed its chromosomal location. Segregating indels in introns were therefore used as markers to map as many *S. vulgaris* orthologues of *S. latifolia* X-linked genes as possible. Segregating ISVS and SNP markers were obtained for six *S. vulgaris* genes, the homologues of the X-linked genes *SlX3*, *CypX*, *SlX7*, *SlX6a*, *DD44X*, and *SlX9*, and scored in the progeny of families SV1 and SV2 (see Methods). Neither mapping family had segregating indels in the introns of *Svcyt*, so the segregation of two SNPs in intron 1 was scored in family SV1 by direct sequencing; informative variants were present in both the maternal and paternal parents. Figure 4.2 shows the consensus map obtained for the two *S. vulgaris* families. The mapped genes fall into a single *S. vulgaris* linkage group, spanning 25cM, except for *Svcyt*, suggesting that this gene moved to the X of *S. latifolia* from another location. We next describe the analysis of this gene in more detail.

4.4.4 *Slcyt* gene structure and function

The complete coding sequence and most of the intron sequences were obtained for *Slcyt.* The gene structure was inferred by comparing the *S. latifolia* cDNA sequence with the genomic sequence (Figure 4.1). BLASTX searches were performed at www.ncbi.nlm.nih.gov to identify homologous genes in other organisms and their functions. The best hit in the *A. thaliana* genome sequence was to a member a family

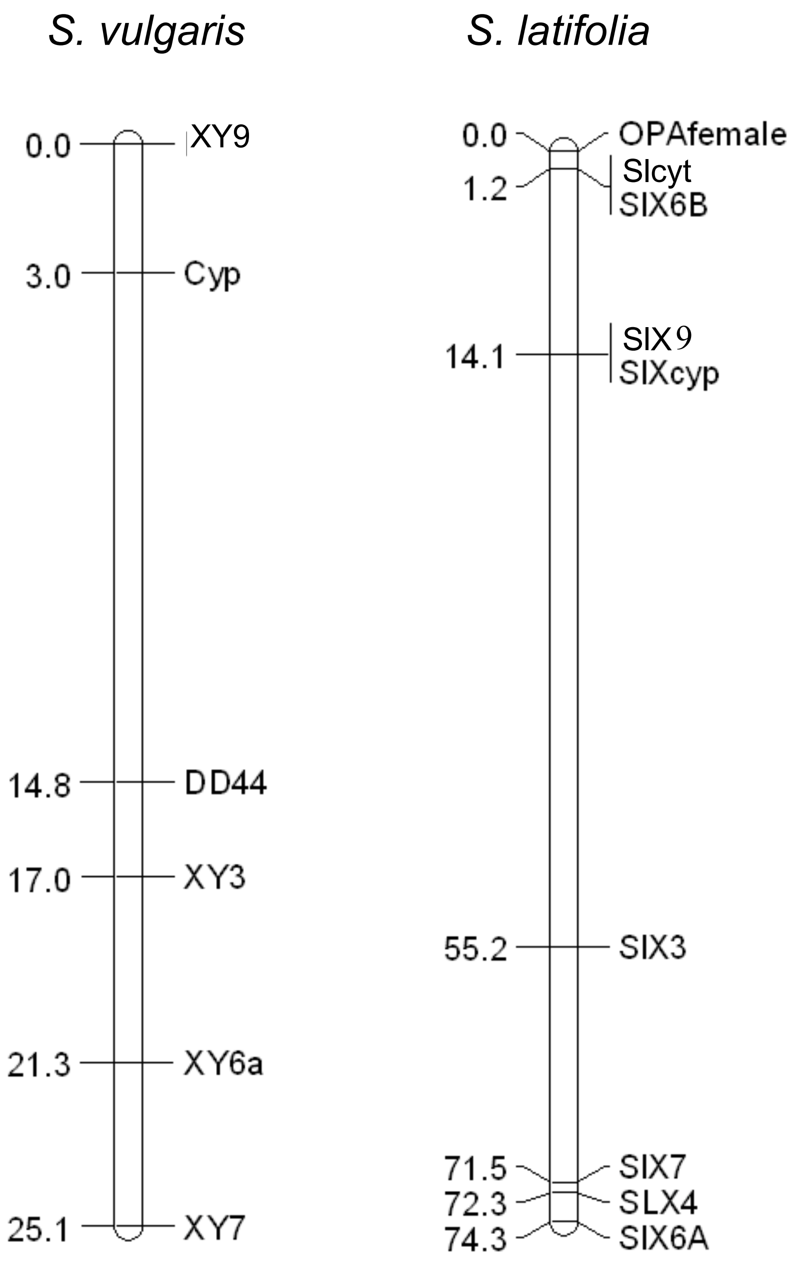


Figure 4.2: Genetic map of X-linked genes in *S. latifolia* and their homologues in *S. vulgaris*. The *cyt* gene falls into a different linkage group in *S. vulgaris*, and is therefore not shown. *DD44X* was mapped onto the *S. latifolia* X chromosome by FILATOV (2005), using a different mapping family.

Table 4.4: Synonymous (K_s) and non-synonymous (K_a) divergence values between *Slcyt* and the sequenes of the orthologous *Cyt* gene in three related species. The lengths of the alignments are given in base pairs (bp).

Species compared	K_S (%)	K_A (%)	Length of sequence (bp)
S. latifolia - S. vulgaris	10	6	273
S. latifolia - S. dioica	1.6	1.3	288
S. latifolia - S. diclinis	1.3	0.4	351

of cytochrome B5 proteins (BLAST e-value 2e^{-41}) that has 61% amino acid identity to *Slcyt*. *Slcyt* also shows significant similarity to cytochrome B proteins in the dicotyledonous plant *Vernicia fordii* (Euphorbiaceae), and in the monocotyledons *Sorghum bicolor* and *Triticum monococcum* (BLAST e-values 1e^{-46} to 4e^{-34}). The homologue of the *Slcyt* gene was sequenced in *S. vulgaris, S. dioica* and *S. diclinis*. Divergence values between *S. latifolia* and the other species are shown in Table 4.4. The *Slcyt* gene has probably remained functional in the species studied. We isolated it from cDNA, showing that the gene is transcribed, and we found no premature stop codons or frameshifts in any of the sequences, and, in comparisons between *S. latifolia* and *S. vulgaris*, K_A is smaller than K_S, suggesting that purifying selection has been acting on the gene.

4.4.5 Searches for a Y-linked copy of *Slcyt* in *S. latifolia* and its close relatives *S. dioica* and *S. diclinis*

No heterozygotes were found among 48 males screened for the microsatellite SlX_STR1 in the *Slcyt* gene, even though we found 19 length variants among these males (Table 4.2), and most females scored were heterozygous (Table 4.3). It therefore seems likely that there is no Y copy. Given that *Slcyt* maps close to the pseudoautosomal region, divergence between the X-linked gene and any Y-linked copy is expected to be low, and conserved primers should amplify a Y-linked copy of *Slcyt*, if present.

To more rigorously test for the possibility that a Y-linked copy of *Slcyt* exists, several approaches were tried. First, primers designed to match regions of conserved amino acid identity between *Slcyt* in *S. latifolia* and the *A. thaliana* or *S. vulgaris* homologues were used to amplify genomic DNA and cDNA from male individuals. No heterozygous SNPs were detected in any of six males tested, four of which were F1 individuals from crosses between different populations. Second, 48 males from 24 different European populations (see above) were sequenced in our diversity analysis, and none carried any SNPs in the region.

We did not obtain any *Slcyt* sequence (not even the X sequence) using an approach in which amplification requires only one primer (either forward or reverse) to match the Y-linked copy (GUTTMAN and CHARLESWORTH 1998). The primers tested were *Slcyt*_ex3_cons_R, *Slcyt*-exon2-R, *Slcyt*_ex1_cons_F, *Slcyt*_F (exon 3), Slcyt_b_R (exon 3), *Slcyt*_ex2_cons_F.

To test whether there is a Y-linked copy in *S. dioica*, we amplified the microsatellite SlX_STR1 in 20 offspring of a *S. dioica* female from a population in Finland, whose genotype at this locus is unknown, using the same primer combination as for *S. latifolia*, and scored the length variants using Genemapper. Among the progeny there were three alleles, 146 bp, 168 bp and 180 bp. Overall, nine females were heterozygous (three 168/180 and six 146/180) and five were 146 bp homozygotes (Table 4.5). In all six males, only one allele was detected (either 146 bp or 180 bp, see Table 4.5). This suggests X linkage and no Y copy (assuming a heterozygous 146/180 maternal parent, which mated with at least three males, carrying 146 bp, 180 bp and 168 bp variants on the X). Autosomal inheritance is not excluded by these results alone (a 146/180 female mated with at least two males). However, it is very unlikely that all six male offspring would be homozygous by chance, but only five out of 14 females (Fisher's exact test, d.f. = 1, $p < 0.05$).

We also genotyped eleven male and six female *S. dioica,* sampled from the wild. All male individuals again each had only one length variant, whereas four females were heterozygotes (Table 4.6).

Table 4.5: Lengths of the region that includes the microsatellite SlX_STR1 in intron 1 of the *Cyt* gene, in twenty offspring of a *S. dioica* female sampled from a wild population (family A2009-1, see text).

Lengths observed (bp)			Numbers of females	Numbers of males
146	168	180		
—	+	+	3	0
+	—	+	6	0
+	—	—	5	3
—	—	+	0	3
Totals			14	6

Table 4.6: Lengths of the region that includes the microsatellite SlX_STR1 in intron 1 of *Cyt* in male and female *S. dioica* individuals from natural populations.

Individual	Location	Sex	Length of amplicon (bp)
99M24.2	unknown	male	129
9-1	unknown	male	146
FS01	Scotland	male	139
FS02	Scotland	male	136
FS03	Scotland	male	133
FS04	Scotland	male	136
FS05	Scotland	male	146
FS06	Scotland	male	139
FS07	Scotland	male	136
FS08	Scotland	male	136
FS09	Scotland	male	133
99K24.1	France	female	129,136
99K22.7	France	female	129,136
FS10	Scotland	female	136,146
FS11	Scotland	female	136,146
FS12	Scotland	female	139
FS13	Scotland	female	139

In *S. diclinis*, however, all three males and three females scored carried two different length variants for *Cyt*. These plants were derived from seeds collected from one female from Spain, and were therefore full- or half-siblings (family A2000-13). Direct sequencing of one male and one female individual from this family (individuals A2000-13-4 and A2000-13-10) confirmed that the PCR products (which span a region of 1,162 sites from exon 1 to within exon 3) were indeed two copies of *Cyt*, although they differ by about 70bp in length due to an insertion in intron 1. There were, however, no SNP variants, which suggests that these are two alleles of the same gene, rather than paralogues. Without a mapping family or population samples from *S. diclinis*, we cannot currently test for sex-linkage in this species, so the two copies could be autosomal or in the pseudoautosomal region of the X and Y. A pseudoautosomal location would result in incomplete sex-linkage of two paternal variants in the F1 of a cross, and in different variants being incompletely associated with the sexes in population samples, so this is potentially testable in the future. It would be interesting to test this, because the *S. diclinis* X is now known to have undergone a translocation of an autosome region (HOWELL *et al.* 2009).

4.4.6 Sequence diversity analyses

To test whether the movement of the *Slcyt* gene to the *S. latifolia* X chromosome was a selected event, we analyzed this sequence further. *S. latifolia* genomic DNA sequences were obtained for a region starting within exon 2, and including the whole of intron 2 and exon 3, and parts of the 3′ UTR, and diversity was analyzed using the European sample of *S. latifolia* males described above. We obtained direct sequences of 45 X-linked copies. The diversity of *Slcyt* was very low: synonymous diversity, π_S, was 0.16% and nonsynonymous diversity, π_A, was 0.07%, based on 237 coding sites. The π_A/π_S ratio is 0.45. The aligned noncoding positions (173 sites) also yielded a low diversity estimate (0.16%).

The HKA test (HUDSON *et al.* 1987) was used to compare *Slcyt* diversity levels with those at four other X-linked genes, *SlX9*, *X7, X4* and *Cyp-X*, by testing the within-species polymorphism of *Slcyt*, after correcting for divergence from their *S. vulgaris* orthologues. The test was significant for the comparisons of *Slcyt* versus *SlX9, X4* and *Cyp-X* (χ^2 = 6.26, p < 0.05; χ^2 = 5.72, p < 0.05; χ^2 = 6.82, p < 0.01, respectively); for the comparison of *X7 vs. Slcyt*, $\chi^2 = 3.57$ and $p < 0.06$, but the number of *X7* sites available was much smaller than for the other three genes (191 bp versus 367, 863, 804 bp, for *SlX9, X4* and *Cyp-X*, respectively), and the power was low. We therefore conclude that the low diversity at *Slcyt* is unlikely to be explained by neutral processes alone, such as a very low mutation rate at the locus.

The low *Slcyt* diversity suggests the possibility of a selective sweep. We therefore performed Tajima's D test, which compares the two estimates of diversity, the nucleotide diversity, π, based on pairwise comparisons of allele sequences, and θ (which is based on the number of segregating sites). At equilibrium under neutrality, the two estimates are expected to be the same, and Tajima's D will be close to zero (TAJIMA 1989), whereas recent directional selection (positive or purifying selection) leads to an excess of low frequency variants, and a negative Tajima's D. The other tests shown in Table 4.7 also detect variants at frequencies differing from the neutral distribution. All the tests performed on *Slcyt* were significant; a skewed frequency spectrum at segregating sites, together with an excess of high-frequency variants which represent the derived state (Fay and Wu's H), suggest the action of positive selection, rather than selection against deleterious variants.

Table 4.7: Tests of neutrality statistics *Slcyt.*

Tajima's D	Fu and Li's D*	Fu and Li's F*	Fu's F	Fay and Wu's H
-1.6	-2.66	-2.72	-2.6	-1.82
$p < 0.05$	$p < 0.01$	$p < 0.01$	$p < 0.05$	$p < 0.01$

4.5 Discussion

The sex chromosomes in *S. latifolia* are believed to have evolved from a single pair of ordinary autosomes, consistent with the fact that all eleven sex-linked genes previously described have Y-linked counterparts (ATANASSOV *et al.* 2001; BERGERO *et al.* 2007; DELICHERE *et al.* 1999; FILATOV 2005; MOORE *et al.* 2003;

NICOLAS *et al.* 2005), and that all four genes previously mapped in *S. vulgaris* (FILATOV 2005) are on a single autosome, as is the newly discovered X-linked gene *SlX9* mapped here. *Slcyt* is the first gene discovered on the *S. latifolia* X that has been acquired from a different genomic location in the course of X-chromosome evolution in this genus, and is now located close to the *SlX6b* gene. In female meiosis, both genes map only about 1cM from the pseudoautosomal marker, OPA, discovered by Di Stilio et al. (1998). Divergence between *SlX6b* and its Y-linked homologue is low (K_S = 4.5%), showing that the X stopped recombining with the Y chromosome only recently at this locus (BERGERO *et al.* 2007). All 46 males from the different European populations contained only a single, hemizygous copy of *Slcyt*. Thus, a translocation was probably involved, which is now fixed within the species. Since no autosomal paralogous copy was detected, the chromosomal segment probably moved onto the X chromosome only (and not the Y also), and was lost from another genomic region at the same time.

It is unknown whether the event affected *Slcyt* alone or whether neighbouring genes were translocated in the same event. *Slcyt* contains introns, showing that it is not a retrogene, so a larger, segmental event is possible, forming a neo-X chromosome. If the event indeed translocated a substantial region into the *S. latifolia* pseudoautosomal region, it might have directly prevented recombination between the X and Y chromosomes in this region (and, once fixed, would not affect recombination between X chromosomes, though it might initially have been deleterious for female fertility in heterozygotes, as is true for many translocations). This event could be similar to that in the medaka fish, where a duplication created a new copy of the autosomal *dmrt1* gene on another autosome, immediately isolating a small region close to the new copy from its homologue (KONDO *et al.* 2006). In medaka, unlike the case of *Slcyt*, the original copy is still present. More information on *S. vulgaris* or *S. latifolia* sequences, including sequencing BAC clones containing the *S. latifolia Slcyt* gene, should allow genes adjacent to *Slcyt* to be found, which would help determine the size of the insertion, and whether any of these genes could have been under sexually antagonistic selection, and could thus have selected for the translocation. Another interesting consequence of movement of a large genome region, containing several essential genes, from an autosome to the X, is that this could explain why YY plants in *S. latifolia* are inviable, since these plants would entirely lack these genes. This alternative to classical genetic degeneration of the Y chromosome has not previously been considered. Such a genome rearrangement would have to be sufficiently advantageous in females to drive the change, however.

Our evidence gives some support for the hypothesis that the translocation was driven by selection. *Cyt* appears to be X-linked in *S. dioica* as well as *S. latifolia*, but not in *S. diclinis*, in which we found two copies in both females and males, so that it is probably still autosomal. Thus *Slcyt* probably moved to the X after the split of *S. latifolia/S. dioica* from *S. diclinis*. This is consistent with a skew in the frequency spectrum still being observable at segregating sites in the *Slcyt* gene. The combination of strongly reduced diversity and a highly skewed frequency spectrum of segregating sites, reflecting a recently reduced effective population size, suggests positive selection when the translocation became fixed (though we cannot exclude the possibility of a selective sweep after its fixation). A recent rapid non-selective fixation could also lead to a uniform haplotype at the locus (TAJIMA 1990). It seems unlikely, however, that such a non-selective event would produce the strongly significant results we find for Tajima's tests and the other tests in Table 4.7, but this possibility should be tested in the future using neutral models.

If selection was involved, it need not have involved the *Slcyt* locus itself. The protein encoded is probably used in the mitochondrial electron transport system. *Slcyt* may therefore be a housekeeping gene, in which case sexual antagonism involving *Slcyt* itself might not be the main driver behind the gene's movement, and one would have to assume selection on another gene translocated in the same event (Jiang et al. 2001; Jin et al. 2001). However, a connection is well-established between mitochondrial electron transport and cytoplasmic male-sterility in plants, possibly due to the high metabolic needs of anther development (CHASE 2007; WARMKE and LEE 1978), and so it is possible that *Slcyt* has some important function in males, and that its loss would benefit females due to sexual antagonism (assuming that its effect is recessive). It will thus be very interesting in the future to compare gene expression of *Slcyt* in males and females and in sex-specific reproductive tissues.

5 High sequence diversity and possible introgression of an X-linked gene on a plant sex chromosome

Contributing authors:

- I carried out the laboratory work, the analysis and writing of the manuscript
- R. Bergero helped with the laboratory work and the analysis
- D. Charlesworth assisted with the analysis and writing of the manuscript
- Plant material was made available from collections of R. Bergero and D. Charlesworth
- R. Bergero handled the greenhouse and DNA extractions and constructed the cDNA libraries
- Helen Borthwick carried out the DNA extractions and assisted in the laboratory
- Sequencing was performed by the Edinburgh University Sequencing service

5.1 Abstract

I describe patterns of DNA sequence diversity in a newly-identified sex linked gene in *Silene latifolia*, *SlX9/SlY9*. The copies on both sex chromosomes appear to be functional, and each maps close to the respective X- and Y-linked copy of another sex-linked gene pair, *SlCyp*. The Y-linked copy has low diversity, similar to what has been found for several other Y-linked genes in *S. latifolia*, and consistent with the theoretical expectations of hitch-hiking processes occurring on a non-recombining chromosome. However, for *SlX9*, we find high diversity (nucleotide diversity for silent sites is estimated to be 4%), much higher than other genes on the *S. latifolia* X chromosome. We evaluate the hypothesis of introgression from the closely related species *S. dioica* as an explanation for the high diversity.

5.2 Introduction

Measuring neutral nucleotide diversity, π, provides information about population size and structure, demographic events, and selection. Estimates of diversity vary within a genome (NACHMAN *et al.* 1998; SACHIDANANDAM *et al.* 2001), either due to regional variation in the mutation rate (HAAG-LIAUTARD *et al.* 2007; WOLFE *et al.* 1989) or different demographic histories at different loci (REICH *et al.* 2002), or both. With recombination and selection acting differently at different sites, loci may differ in their age, i.e. in the expected time, $t = 2N_e$, back to the most common recent ancestor; positive selection reduces t (and hence $\pi = 4\ N_e\mu$), whereas balancing selection increases it.

In contrast to the situation for autosomal and X-linked loci, on a non-recombining chromosome, such as the Y chromosome, all sites are completely linked and experience the same effective population size; hence we expect diversity at Y-linked loci to be much more homogeneous, differences in π mainly reflecting differences in μ. Because of selective interference effects among linked loci, such as weak selection Hill-Robertson interference (COMERON 2008; HILL and ROBERTSON 1966; MCVEAN and CHARLESWORTH 2000), genetic hitchhiking due to positive selection (KAPLAN *et al.* 1989; MAYNARD SMITH and HAIGH 1974) or the elimination of deleterious variants (background selection and Muller's ratchet (CHARLESWORTH *et al.* 1993; GORDO *et al.* 2002; MULLER 1964), we expect diversity values to be lower on the Y compared to the X or autosomes, taking into account the smaller number of Y chromosomes in the population.

Silene latifolia is a dioecious plant and a model system for the evolution of young sex chromosomes. In this system, synonymous diversity values of six X-linked genes studied (*SlX1, SlX4, DD44-X, SlssX, Sl-Cyt, Cl-Cyp*) vary between 0.07% and 5.1% (ATANASSOV *et al.* 2001; BERGERO *et al.* 2008; LAPORTE *et al.* 2005; Bergero, unpublished data; Chapter 4) (see Table 5-1), the lowest value for *SlssX* apparently being due to a recent selective sweep in the genomic region that did not affect diversity at the nearby *DD44-X* (FILATOV 2008). We do not have estimates for recombination rates per physical distance for the Silene sex chromosome, but a rough estimate suggest 18cM per MB of sequence. Four Y-linked genes studied (*SlY1, SlY4, DD44-Y, SLAP3Y*) have silent diversity values between zero and 0.28% (ATANASSOV *et al.* 2001; LAPORTE *et al.* 2005; MATSUNAGA *et al.* 2003) (see Table 5-1), which is consistent with the idea that interference among selected sites reduces N_e on the Y chromosome.

Apart from variable selection pressures and mutation rates among loci, introgression might play a role in shaping within-species polymorphism, by increasing the coalescent time for a given locus. *S. latifolia* forms natural hybrids with its closely related sister species *Silene dioica* (DESFEUX *et al.* 1996; MINDER *et al.* 2007; MINDER and WIDMER 2008), but the two species are geographically and ecologically distinct, with *S. latifolia* having white flowers, a generally wider distribution and growing in dry, open habitats, whereas *S. dioica* has red flowers, and is found mainly in Northern Europe, at the margins of woodlands (BAKER 1947; 1948; KARRENBERG and FAVRE 2008). Also the type of pollinators differ, with *S. latifolia* being mainly visited by the moth *Hadena bicruris*, and *S. dioica* during the day by bumblebees and butterflies (BOPP and GOTTSBERGER 2004; MINDER *et al.* 2007). However, data from three known sex-linked genes, *DD44*, *SlX1* and *SlX4* (IRONSIDE and FILATOV 2005; LAPORTE *et al.* 2005), as well as AFLP markers (KARRENBERG and FAVRE 2008;

MINDER *et al.* 2007; MINDER and WIDMER 2008), suggest that introgression of *S. dioica* sequences into *S. latifolia* is common in nature. In line with this, hybrids carrying pink flowers can be seen frequently in the wild.

In order to understand sex chromosome evolution in its early stages, I isolated more genes from the *S. latifolia* X and Y chromosome, to test for Y chromosome degeneration and the effects of different chromosomal environments on nucleotide diversity. Here, we describe a newly identified sex-linked gene pair in *S. latifolia*, *SlX9*/*SlY9*, and analyse it with respect to diversity levels, Y chromosome degeneration and introgression. We find that synonymous (but not non-synonymous) diversity levels at *SlX9* are highly elevated, suggesting that diversity varies across more than two orders of magnitude across the *S. latifolia* X chromosome. High diversity at *SlX9* is most likely due to introgression from *S. dioica*, which might be confined to X-linked and autosomal loci only. In contrast, the Y linked copy, *SlY9*, which seems to be functional, has very low diversity levels. We do not find any evidence for coding sequence degeneration of *SlY9*, but we find that the Y has accumulated additional intronic sequence, as observed previously for the Y-linked genes *DD44-Y* and *SlY3* (MARAIS *et al.* 2008) or *SlCyp-Y* (BERGERO and CHARLESWORTH 2009).

Table 5-1: Levels of diversity, π (%), in previously studied sex-linked genes of *S. latifolia.*

Gene	**Synonymous sites**	**Non-synonymous sites**	**Reference**	Gene	**Synonymous sites**	**Non-synonymous sites**	**Reference**
SlX1	2.3 or 2.1	0.24	(ATANASSOV *et al.* 2001; BERGERO *et al.* 2008)	*SlY1*	0	0.006	(ATANASSOV *et al.* 2001)
SlX4	5.1or 4.4	0.592/0.45	(BERGERO *et al.* 2008; LAPORTE *et al.* 2005)	*SlY4*	0.000	0.000	(LAPORTE *et al.* 2005)
DD44X	2.4	0.44	(BERGERO *et al.* 2008; LAPORTE *et al.* 2005)	*DD44Y*	0.277	0.000	(LAPORTE *et al.* 2005)
SlSSx	0.07	0.03	(BERGERO *et al.* 2008)	*SLAP3Y*	0.12	0.083	(MATSUNAGA *et al.* 2003)
SlCyt	0.17	0.08	see Chapter 4				
CypX	1.098	0.020	R. Bergero, unpublished data				

5.3 Methods

5.3.1 Plant materials

Sex-linkage of *SlX9* was established using the mapping family H2005-1 (BERGERO *et al.* 2007), which is a full-sib cross between F1 offspring whose parents came from different European populations (male E2004-17-1, from the Netherlands, and female E2004-11-1, from Canche, Northern France). 92 plants from this family were used to map its location on the X chromosome map. The mother of the mapping family is a heterozygote for two X-linked alleles that produced PCR products of different lengths (bands of about 450 and 600 bp, see Figure 5-1), which were used for genetic mapping, as described below. The panel of 38 Y-deletion mutants used to find the location of *SlY9* is described in BERGERO *et al.* (2008). To study sequence diversity, we used a sample of 46 males from 24 different European populations, covering the entire range of the species, and six *S. dioica* individuals, including plants from France and Finland (listed in Table 5-2 and 5-3).

5.3.2 PCR amplifications and primers

SlX9 was identified from a *S. latifolia* cDNA library derived from male leaf primordia, and shown to be a sex-linked gene (see Chapter 4). The complete cDNA sequence was obtained, and primers were designed based on this sequence. As described below, it proved very difficult to sequence this gene in its entirety from all our sampled individuals, and only partial genomic sequences were obtained, with different regions sequenced from different species, and from the X and Y copies (see Figure 5-2 and Table 5-4 below). To obtain sequences, new primers were designed from the sequences yielded by the initial primers; these are listed in Table 5-5.

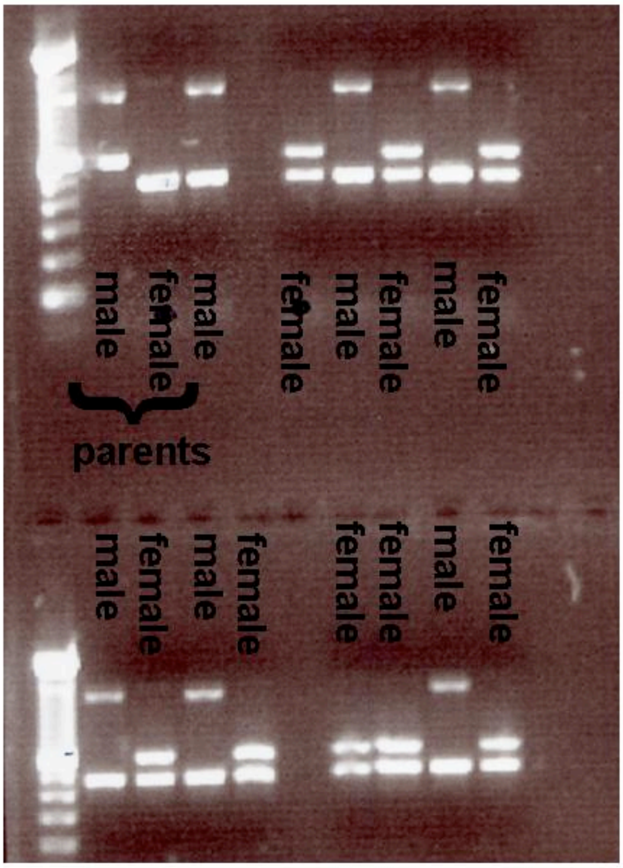


Figure 5-1: PCR products of *SlX9* in family H2005-1. The female parent carries only one variant (*SlX9*) and the father two differently sized variants (*SlX9* and *SlY9*). All female offspring inherit the two *SlX9* copies, whereas all sons inherit the *SlY9* copy from the father as well as the *SlX9* variant from the mother.

Table 5-2: Origins of male individuals for the *SlX9/SlY9* diversity study, the intron type of the X-linked copy ("long" or "short") and the broad geographical region.

* **Geographical regions:** 1 = "Northern Europe"; 2 = "North-Eastern Europe"; 3 = "Mediterranean group"; 4 = "Spain and Portugal"

ID	Country	Population	Intron type	Geographical region*
1	Germany	Bissendorf	short	1
2			short	1
3	Sweden	Oland, Grasgard	short	1
4			long	1
5	France	Relais des Chenes	-	3
6			short	3
7	Greece	Ioanninon	short	3
8			long	3
9	Poland	Krakow, near Conference Centre	long	2
10			short	2
11	Denmark	Aarhus Botanic Garden, origin unknown	short	1
12			short	1
13	Austria	Vienna, Heiligenstadt	short	2
14			short	2
15	Norway	Lardalsoyri	short	1
16			short	1
17	Greece	Evrou	short	3
18			short	3
19	Italy	near Tarquinia, Lazio	-	3
20			long	3
21	Estonia	Vinupea	long	2
22			-	2
23	Austria	Dietmans	-	2
24			long	2
25	France	Canche	short	1

Table 5-2, continued

26	France	Canche	short	1
27	Portugal	Serre de Nogere	long	4
28			long	4
29	Netherlands	River Kraal	long	1
30			long	1
31	France	Vitry en Artois	short	1
32			short	1
33	Italy	Piedmont, Ceres	-	3
34			short	3
35	UK	Dalkeith	long	1
36			long	1
37	UK	Somerset, Burnham-on-Sea	short	1
38			short	1
41	Portugal	Segier-Chavez	long	4
42			long	4
43	Spain	Madrid, San Lorenzo de El Escorial	short	4
44			short	4
45	Ukraine	Ukraine	short	2
46		Ukraine/Belarus border	short	2
47	Germany	near Glaubitz	short	1
48			short	1

Table 5-3: *S. dioica* individuals used in this study

ID	Location	Sex
99M24.2	unknown	male
99M9.1	unknown	male
99K24.1	France	female
A2009_1_female_1	Finland	female
A2009_1_male_7	Finland	male
A2009_1_male_9	Finland	male
A2009_2_female_1	Finland	female

Table 5-4: Regions amplified for the diversity study

Gene	Sequence for diversity study	No. coding sites analyzed	No. noncoding sites analyzed
SlX9	intron 1 (partial), exon 2, intron 2, exon3, intron 3 (partial)	255	178
SlY9	intron 2 (partial), exon 3, intron 3, exon 4(partial)	270	222

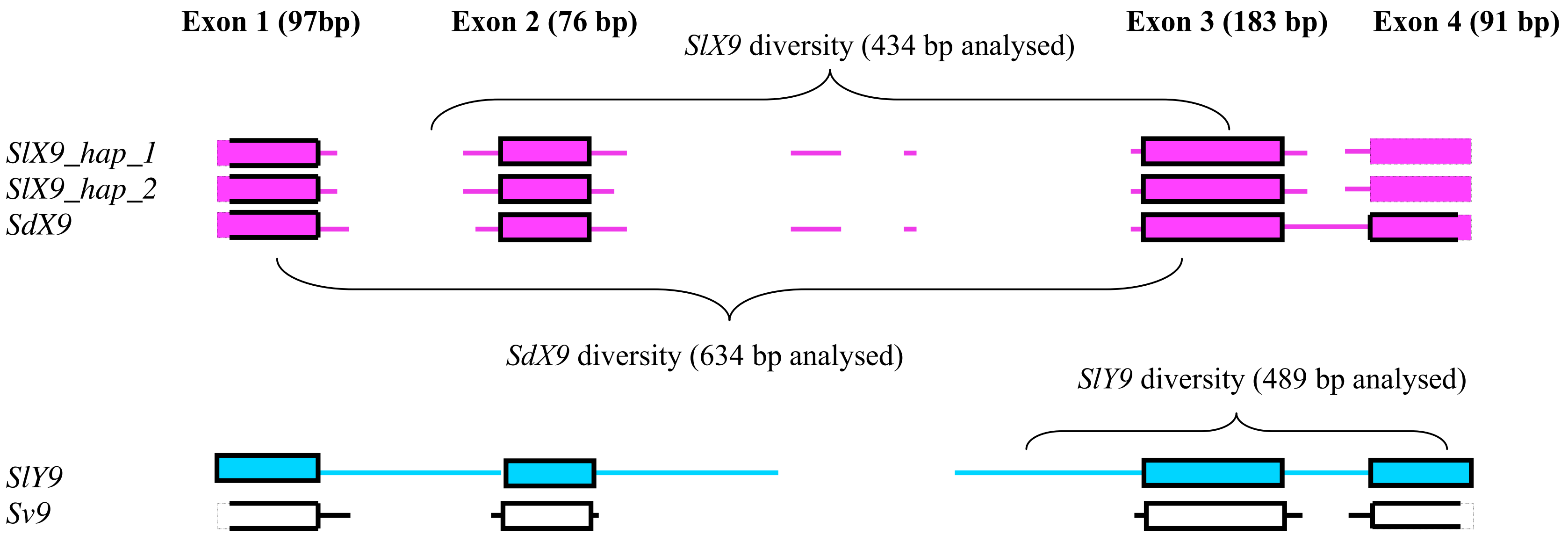

Exon 1 (97bp)
Exon 2 (76 bp)
Exon 3 (183 bp)
Exon 4 (91 bp)
SlX9 diversity (434 bp analysed)
SlX9_hap_1
SlX9_hap_2
SdX9
SdX9 diversity (634 bp analysed)
SlY9 diversity (489 bp analysed)
SlY9
Sv9

Figure 5-2: Schematic view of the alignment of *SlX9, SdX9, SlY9* and *Sv9*. Sizes of introns and exons are not drawn to scale. Both X-linked intronic variants are shown (*SlX9*_hap_1 and *SlX9*_hap_2). Thick lines around exons: sequences available for at least one individual. (The original *S. latifolia* cDNA clone from which the gene was identified contained the whole open reading frame; sizes of exons for which we do not have complete sequences are drawn based on the assumption that exon sizes are the same as for the cDNA.) Note: Sequences used in the diversity studies of *SlX9*, *SdX9* and *SlY9* only cover parts of the gene sequences, as indicated.

Table 5-5: Primers used in this study

Name	Sequence	Notes
RB18_F	CTTGTGGAACTTCTGGTGGAAG	establishing sex-linkage;
RB18_R	GTCCAATCACATTCAAGTCTCTCC	deletion mapping of *SlY9*
RB18_male_intron_2_F	TCTTTCACACCCAATTTGATCC	amplify *SlY9* with RB18_E3_rev
RB18_male_intron_2_R	GTACAGGGAAGAGCAAAGCAC	
RB18_exon_1_F	AGCTAGCAGTTTTGCAGCATC	amplify *SlY9*
RB18_Y_E3-R	GACATTCAAGTCTCTCCTCAGCCAA	amplification of cDNA to get Y copy
RB18-Intron2-male-F-2	AAGGACAACAATTCAATGGGATG	
RB18-Intron2-male-F-3	GGGATGGAGGGAGTATGTTATTATTG	*SlY9*-diverstiy with RB18-3'UTR
RB18-3'UTR	GATGAATCTAAAATCAAACAGTGAAAC	
RB18_exon_1_F	AGCTAGCAGTTTTGCAGCATC	
RB18_exon_4_R	TCTTCAGTCCTTCCTTTGAAGC	*SlX9*-diverstiy
RB18-male-exon1-F	ACTCTCTCTCGCTCTTACTCC	
RB18_male_exon3-R	GACATTCAAGTCTCTCCTCAGCCAA	
RB18-E3-R-beg	TCCTCAGCCAGTCTCTTTGAA	
RB18-E2-F-beg	GTTGCTGCAGTGAACCCTCT	
RB18-EXON2-R-Male	TGCCTCAGAGGGTTTACTGC	

PCR amplification was generally done using Taq JumpStart^{TM} (Sigma-Aldrich), and the following conditions: initial denaturation at 95℃ for 5min, 10 cycles of denaturation at 95℃ (30 sec), annealing at 55-58℃ (30 sec), extension at 72℃ (1-1.5 min), final extension at 72℃ for 15 min. PCR amplicons were cleaned using ExoSAP-IT (Amersham Biosciences, Tokyo) and sequenced on an ABI 3730 capillary sequencer (Applied Biosystems) and sequences edited using Sequencher 4.7.

The two X-linked alleles in the mapping family (see above) were cloned from PCR products and sequenced. Primers used to amplify across introns 1, 2 and 3 are listed in Table 5-5. The gene structure of *SlX9* was then inferred by comparing the *SlX9* genomic sequence with its cDNA sequence, as well as by comparisons with *A. thaliana* and *S. vulgaris* gene structures. A BLASTN search was performed at www.ncbi.nlm.nih.gov to identify homologous genes in other organisms and their functions.

5.3.3 Obtaining and mapping the Y-linked homologue

As described in Chapter 4, using primers RB18_F and RB18_R (Table 5-5) to amplify DNA from male and female plants from family H2005-1, yielded a male-specific PCR product of about 1.2kb. To sequence the Y-linked homologue of *SlX9*, the longer, male-specific allele (see Figure 5-1) was cut from the agarose gel, cloned and sequenced. To obtain the 5′ coding sequence of *SlY9*, which was not present in the region initially sequenced, the cloned PCR products from family H2005-1 were also used to design new, Y-specific primers from this sequence (Table 5-5). These primers were also used for the *SlY9* diversity study (see below).

To test whether the Y-linked gene was expressed, we did nested PCR using cDNA derived from male flower tissue. The first round of PCR amplification used the primers TsShort and RB18_R (Table 5-5), where TsShort matched the sequence to which the cDNA was ligated. The second round used TsShort and the Y-specific primer RB18_Y_E3-R (Table 5-5).

For deletion mapping *SlY9* on the Y chromosome, PCR amplifications were scored in the deletion mutants (see above) using the primers RB18_F and RB18-R (Table 5-5). The location of *SlY9* was inferred by comparing the presence/absence of the Y-linked, larger, fragment to the presence/absence of other Y-linked genes (BERGERO *et al.* 2008; ZLUVOVA J *et al.* 2007; ZLUVOVA *et al.* 2005).

5.3.4 Diversity of *SlX9* and *SlY9* and linkage disequilibrium analysis

To study sequence diversity of *SlX9* and *SlY9*, parts of the genes (listed in Table 5-4) were amplified in the 46 male individuals from different European populations (Table 5-2); The X-linked copy was amplified from each plant using the primers RB18_exon_1_F and RB18_exon_4_R, and the Y-linked copy with either RB18-Intron2-male-F-2 and RB18-3´UTR, or RB18-Intron2-male-F-3 and RB18-3´UTR respectively (Table 5-5). The sequence fragments were assembled and aligned manually using the program Se-Al v.2.0a11 (Se-Al: Sequence Alignment Editor, http://evolve.zoo.ox.ac.uk/).

Sequence diversity was analyzed using DnaSP 4.0 software (ROZAS and ROZAS 1999), excluding indel polymorphisms. To estimate between-population differentiation, the *SlX9* and *SlY9* sequences were divided into four broad geographic groups based on their location of origin ("Northern Europe", "North-Eastern Europe", "Mediterranean group", and "Spain and Portugal", see Table 5-2), K_{ST} statistics were computed in DnaSP. A NJ tree of the X-linked sequences was constructed in MEGA 3.1 (KUMAR *et al.* 2004).

To test for introgression from the sister species *S. dioica*, we used the principle that introgression will cause variants from one species to be more often found in the same haplotype than expected based on their frequencies in the hybrid population, i.e. there will be positive linkage disequilibrium among segregating sites, specifically in regions containing multiple successive variants from *S. dioica*. The associations are expected to last until they are broken up by recombination events, so that the physical distance over which we observe positive LD can be used as a measure of the time when hybridization occurred and/or the strength of selection against hybrid individuals. We used DnaSP 4.0 to calculate D´, a measure of linkage disequilibrium among segregating sites, standardized relative to its maximum possible value.

5.3.5 Tests of neutrality and recombination estimates

Using these population samples of *SlX9* and *SlY9*, several neutrality tests were performed in DnaSP 4.0, including Tajima's D, Fu and Li's D* and F*, Fu's F and Fay and Wu's H statistics (FAY and WU 2000; FU 1997; FU and LI 1993; TAJIMA

1989). Levels of statistical significance were estimated using coalescent simulations in DnaSP 4.0, conservatively assuming no recombination.

We also used DnaSP 4.0 to calculate an estimate of the recombination parameter, $R = 3N_er$ (for *SlX9*); the minimum number of recombination events; Strobeck's *S* statistic, which gives the probability of sampling the same or smaller number of haplotypes as observed in the population sample, given an estimate of θ (STROBECK 1987).

5.3.6 Comparisons with outgroup species

Using the primer pair RB18_exon_1_F and RB18_exon_4_R (Table 5-5), the homologue of *SlX9/SlY9* was amplified in *S. vulgaris*, a gynodioecious species that lacks sex chromosome and forms an outgroup to the *S. latifolia/S.dioica* clade.

Amplification in *S. dioica* was done using different combinations of primers, listed in Table 5-5 (RB18-male-exon1-F and RB18_male_exon3-R; RB18_exon_1_F and RB18-E3-R-beg; RB18-E2-F-beg and RB18_exon_4_R; RB18-male-exon1-F and RB18-EXON2-R-Male; RB18-male-exon1-F and RB18_exon_4_R). We call these homologues *Sv9* (for *S. vulgaris*) and *SdX9* (for *S. dioica*); we infer that *SdX9* is X-linked, but we did not obtain the Y-linked homologue for this species (see Results). The HKA test (HUDSON *et al.* 1987), as implemented in DnaSP 4.0, was used to compare the diversity levels at *SlX9* and *SlY9*, taking into account the different ploidy levels, and using *Sv9* as the outgroup sequence.

To test whether *SlY9* has an accelerated rate of mutation, all fourfold degenerate sites were extracted from the *SlX9/Y9* coding sequence using DnaSP 4.0. The baseml program of PAML was used to compare the rates of evolution along the 4 branches of the phylogenetic tree, using *Sv9* as the out group sequence: a model that assumes a single rate of evolution for all branches ("clock = 1") was compared with a model that assumes a different rate for the Y-linked branch compared to the X-linked and autosomal genes ("clock = 2"). Because there were only 72 fourfold degenerate sites, we also performed the same test using all 444 coding sites of the alignment.

The codeml program of PAML was used to estimate K_A/K_S ratios along the branches leading to *SlX9* and *SlY9* respectively, again using *S. vulgaris* as an outgroup, and allowing each branch of the tree to have its own rate of evolution. The likelihood of obtaining the data under a model in which all three branches of the tree have the same K_A/K_S ("model = 0") was compared with a model in which there was one K_A/K_S ratio in the branch leading to *SlY9*, and one for the branches leading to *SlX9* and *Sv9* respectively ("model = 2").

Divergence between the X- and Y-linked copies of *SlX9-Y9* was estimated using DnaSP 4.0. The exonic sequence of male E2004-15-1 (from Serre de Nogere, Portugal) was compared to the set of X-linked sequences amplified for the diversity study (255 coding sites). Synonymous and nonsynonymous divergence values were calculated using DnaSP 4.0.

In *S. latifolia*, sex-linked genes differ in their degree of X-Y divergence, suggesting that recombination initially stopped in the region of the sex chromosomes where female suppressor/male sterility genes evolved (CHARLESWORTH and CHARLESWORTH 1978), and was suppressed between other X-Y homologues at later stages. We used the divergence between *SlX9* and *SlY9* to test whether the significant correlation between the genetical map position of a gene on the X and its divergence from its Y-linked homologue still existed when *SlX9* was added to the dataset of BERGERO *et al.* (2007). If recombination has been gradually repressed along the X-Y axis, one would expect to see such a correlation.

5.4 Results

5.4.1 Discovery of the new gene

As outlined in Chapter 4, our segregation results for a gene amplified using the primer pair RB18_F and RB18-R (see Methods and Table 5-5 above) suggested that the sequence corresponds to a sex-linked gene, which was named *SlXY9*. Figure 5-1 shows the segregation of the informative intron size variant in the family H2005-1. A longer genomic DNA fragment is inherited by all male offspring (indicating Y-linkage), whereas the daughters inherited one copy from the mother and one from the father.

The *SlX9* cDNA sequence contains a continuous open reading frame of 444 bp coding sites (148 amino acids). BLAST tests showed that the gene is similar to the photosystem I subunit of *A. thaliana* (e-value 2e-37 with GeneID: 837358, TAIR:AT1G08380) and to undefined membrane proteins of tobacco, wheat and rice (BLAST e-values 1e-38 to 9e-37). It is thus probably a housekeeping gene.

Partial genomic sequences of *SlX9* and *SlY9* were obtained from males E2004-15-1, E2004-17-1 and E2004-1-9 (whose Y haplotype came from the pollen donor 98E-6/9 (USA)). The results show that the original cDNA corresponds to the X-linked copy (its JC-corrected synonymous divergence from the set of X-linked sequence of the diversity study is 8.8%, based on 61 synonymous sites, similar to the within-species synonymous diversity of *SlX9*, see below). The cDNA sequence from male flower tissue overlapped with the exonic *SlY9* sequence of male E2004-15-1 by 94bp, and the sequences in this region were identical, whereas there were 6 sites with differences from all X-linked sequences, including those obtained in our diversity study (see below). Retrieving a Y-linked copy from cDNA suggests that *SlY9* is expressed. The mean synonymous divergence of all Y sequences from the cDNA was only slightly lower than that for the X (8.6% with JC correction), but there are fixed differences between the X and Y sequences.

Comparisons between the original cDNA clone and all genomic sequences of *SlX9* and *SlY9* that we could obtain (including those obtained in our diversity study, see below) suggest that there are 4 exons in *S. latifolia* (Figure 5-1), implying the presence of one intron more than in the *A. thaliana* putative homologue. These comparisons revealed that the distinctive large Y-linked band identified in family H2005-1 is due to presence of extra sequence in intron 2 of theY-linked copy. The complete intron 2 was obtained only for one male plant (E2004-17-1, the father of mapping family H2005-1); for the diversity study, the forward primer was located within intron 2, so that the full length of the intron in other male plants is not known (the length of sequence obtained for intron 2 in other plants was only 86bp or less). No BlastN or BlastX matches were found for the insertion in *SlY9*, and no repetitive sequences were detected using the RepeatMasker program (www.repeatmasker.org). However, we cannot exclude the possibility that the insertion was caused by a TE, since its sequence might represent a new TE type, or could have changed too much to be recognizable as a known type.

5.4.2 X and Y haplotypes

To study sequence diversity, we obtained 40 X and 46 Y sequences from a total of 46 males from different European locations (amplified using primer pairs listed in Table 5-5: for *SlX9*, the primers RB18_exon_1_F and RB18_exon_4_R were used, and for *SlY9* they were RB18-Intron2-male-F-2 and 3' UTR-R, or RB18-Intron2-male-F-3 and 3' UTR-R). The PCR amplifications always yielded just one copy, which was in all cases clearly identifiable as either *SlX9* or *SlY9*, using the intron length variant that distinguishes the Y-linked alleles (see above). The gene is therefore present as single copy in the genome, and is sex-linked throughout the species' range. No frame shift mutations or premature stop-codons were found in the coding regions in any of the *SlY9* (or *SlX9*) sequences.

Among the X-linked sequences, we found two distinct sequence types. In thirteen *SlX9* sequences, intron 2 was ~ 485bp, whereas in the other 27 it was only ~ 380bp. The intron sequences of the two types were highly diverged, and were aligned manually (see Methods). We discuss these two sequence types further below.

All three introns of *SlY9* are longer than those of *SlX9* or the *S. vulgaris* homologue (*Sv9* in Figure 5-2), and the complete intron 2 from the genomic sequence of male E2004-17-1 is about 600 bp longer than in the longer *SlX9* type just described (see also Figure 5-1). This suggests, by parsimony, that the intron sizes have expanded in the Y-linked copy, consistent with previous findings of non-coding sequence accumulation on the Silene Y chromosome (CERMAK *et al.* 2008; HOBZA *et al.* 2006), and longer introns for the Y-linked genes *DD44Y* and *SlX3* (MARAIS *et al.* 2008), and the observed expansion of the *Drosophila miranda* neo-Y and the non-recombining Y-like region of papaya (BACHTROG *et al.* 2008; LIU *et al.* 2004).

5.4.3 Location of the *SlY9* gene on the X and Y chromosomes, and X-Y divergence

As described in Chapter 4, in the genetic map of the X chromosome, the X-linked *SlX9* gene is closely linked to *SlCyp-X*, a previously described sex-linked gene

(BERGERO *et al.* 2007). To test whether the Y-linked copies of these two genes are also close, we used deletion mapping (see Methods). This showed that *SlY9* is always co-deleted with *SlCyp-Y*, suggesting that these genes have been physically close since recombination stopped in the relevant region of the sex chromosomes.

Because only parts of the sequences could be obtained, estimating divergence between types of sequence is difficult. To estimate X-Y divergence, we compared 392 coding sites of the longer sequence from male E2004-15-1with the original cDNA clone (which, as shown above, is an X-linked sequence); this yielded a silent site divergence estimate (K_S) of 15.3%, after Jukes-Cantor (JC) correction for saturation, and non-synonymous divergence (K_A) was 0.069%. Similarly, (JC-corrected) K_S was 14.4% and K_A 0.026% when 255 nucleotides in exons 2 and 3 from the *SlX9* diversity study and *SlY9* from male E2004-15-1 were compared. There were no fixed non-synonymous differences, i.e. all nonsynonymous differences between *SlX9* and *SlY9* were polymorphisms in the *SlX9* sequences. Similar estimates were obtained whether we include all X sequences, or only the longer or shorter type. Because *SlX9* has high sequence diversity (see below), we also estimated the net divergence between *SlX9* and *SlY9* using these sequences (subtracting the average of the diversity values); this yielded 6.25% for silent sites and 9.8% for synonymous sites (based on 255 coding sites). Because *SlX9* maps to the same position on the X chromosome as *SlCyp-X*, both genes should have stopped recombining at the same time, and they should thus have similar sequence divergence. The synonymous divergence between *SlCyp-X* and *SlCyp*-Y is estimated to be 6.1% (BERGERO *et al.* 2007), slightly lower than for *SlX9* and *SlY9*. This difference may simply be inaccuracy due to the small numbers of sites in our *SlXY9* alignments. The relationship previously found between the position on the genetic map of the X and its X-Y divergence (BERGERO *et al.* 2007) remains significant when *SlX9* is added to the dataset (Figure 5-3; p-value for the regression < 0.1%).

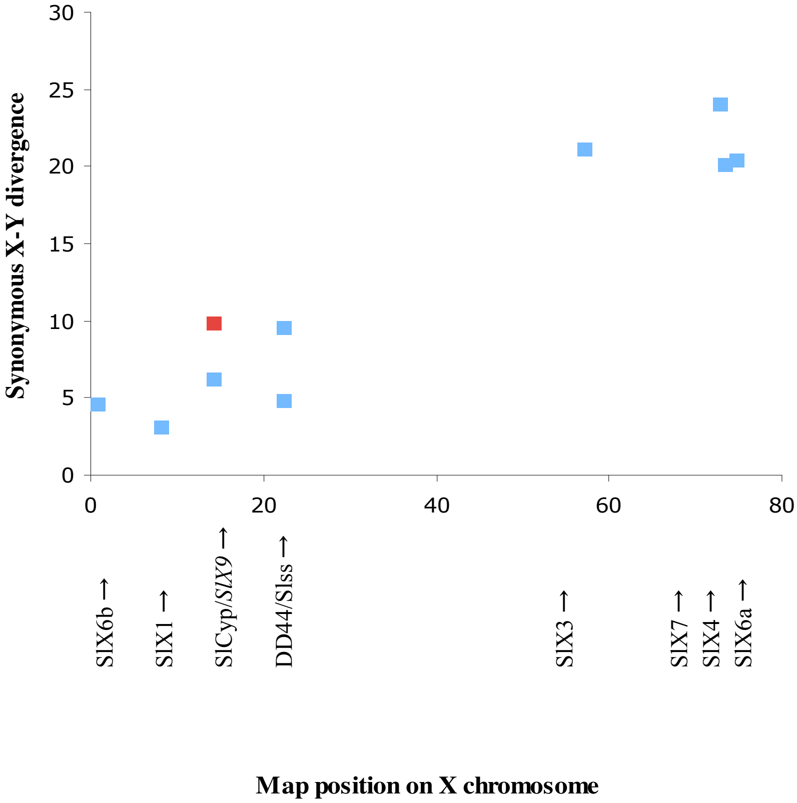


Figure 5-3: Synonymous divergence values between sex-linked genes are plotted against the map position of the respective X-linked gene. *SlX9* is shown as a red square, all other genes as blue squares.

5.4.4 Divergence from outgroup species

Several X-Y pairs of genes have been found to have higher mutation rates in the Y copies ((FILATOV 2005; FILATOV and CHARLESWORTH 2002)), and we tested the *SlX9* and *SlY9* genes for this difference. Both *SlX9* and *SlY9* have similar divergence from the *S. vulgaris* homologue (JC-corrected distances for *SlX9 vs. Sv9* were 14.6% for 90 silent sites, and K_S = 21% and K_A = 0.5%, based on 255 coding sites; for the Y, K_S = 21% and K_A= 0, based on 261 coding sites). PAML analysis confirms that *SlY9* does not evolve significantly faster than the X-linked copy, considering only 4-fold degenerate sites or all coding sites (Table 5-6). Furthermore, the d_N/d_S ratio on the branch leading to *SlY9* does not differ significantly from that on the other branches (Table 5-6). The overall d_N/d_S estimate is 0.077. Together with the lack of frame shift mutations or premature stop-codons in *SlY9*, as well as its expression as mRNA, these results suggest that the *SlY9* gene is still functional.

Table 5-6: PAML-based tests of rate differences between *SlX9* and *SlY9*, using *Sv9* as an out group sequence.

Program	Comparison	No. of sites	model	χ^2	d.f.	p
baseml	4fold degen. sites	72	clock = 1 clock = 2	2.67	1	0.102
baseml	all coding sites	444	clock = 1 clock = 2	0.32	1	0.57
codeml	K_A/K_S	438	model = 0 model = 2	1.51	1	0.22

5.4.5 Sequence diversity in *S. latifolia*

We sequenced a portion of *SlX9* starting within intron 1. Within our *S. latifolia* sample, synonymous diversity of the X-linked copy, *SlX9*, was π_S = 9.2% with JC correction, which is very high, and greatly exceeds the values for other X-linked genes (see above); non-synonymous diversity, π_A, was only 0.05%. These estimates are based on only 85 alignable codons in exons 2 and 3. Silent site diversity is only half the above value (4.0% with JC correction, based on 240 sites), still a high value. As described further below, the high diversity results, at least partly, from the presence of two *SlX9* sequence types (the two X haplotypes with different intron sizes described above, see Figure 5-2). Silent site diversity with JC correction within each set of X-linked sequences was 3.2% in either case, suggesting that the net silent site divergence between the two groups is about 1.3%.

The Y-linked copy, *SlY9*, has substantially lower diversity. In 270 sites of coding sequence, not a single polymorphism was detected among our 46 Y sequences from different European populations, and there were only two SNPs in the introns. For silent sites (283 sites), nucleotide diversity was 0.163% representing a X/Y ratio of silent diversity of 24.5, using the estimate for the X chromosome above, though it should be noted that the X and Y estimates of π are based on different gene regions (see Figure 5-2), and that the numbers of sites analyzed in either set of sequences is small. There was also one indel polymorphism in intron 3 of *SlY9*: the sequences from two males from a population in Greece had an insertion of the triplet TCA. The diversity at *SlY9* is significantly lower than in *SlX9* by an HKA test, taking ploidy levels into account (all sites: $\chi^2 = 5.5$, $p < 0.05$).

Given that the *SlY9* diversity is lower than expected purely from the lower Y effective population size, the reduction is probably due to hitch-hiking processes in the non-recombining regions of the Y chromosome, which include most of this chromosome. In *S. latifolia*, there were three fixed differences between the X- and Y-linked copies (based on 231 bp of aligned sequence), suggesting that the gene has been in the non-recombining region of the Y for some time.

SlY9 and *SlX9* both showed significant differentiation between populations (see Methods), with the *SlX9* K_{ST} estimate being 0.109 ($p < 0.001$) and the estimate for *SlY9* = 0.122 ($p < 0.05$), but the latter is highly inaccurate, being based on only 2 segregating sites.

5.4.6 Introgression from *S. dioica*?

The high *SlX9* diversity is puzzling. Given the genetic evidence mentioned above that there are no duplicate copies of these genes, we tested whether the presence of two *SlX9* sequence types described in the preceding section could be explained by introgression of the *S. dioica* orthologue of *SlX9***.** Very recent introgression seems unlikely, because the short length type was found in populations from the Mediterranean region, where *S. dioica* is absent (PRENTICE *et al.* 2008). Both intron sizes were found within *S. latifolia* populations from Sweden, Poland (where *S. dioica* is also present), and in Greece, and the short *SlX9* sequence was also found in Italy and Spain.

To test this further, we sequenced portions of the gene from *S. dioica* (*SdX9* in Figure 5-2). All four *S. dioica* males only had one copy, whereas one of the two females contained heterozygous SNPs, so it appears that only the X-linked copy amplified from *S. dioica*. In support of this conclusion, mean silent site divergence between *SdX9* and *SlY9* was 8.3% (but based on only 54 sites), higher than the average (raw) divergence between *SlX9* and *SdX9* (which was 2.7%, similar to estimates from other genes in these species, which range from a net silent site divergence of 1.1% for *SlX1* (ATANASSOV *et al.* 2001) to a synonymous divergence of 4.4% for *Sl-Cyp* (BERGERO and CHARLESWORTH 2009)). Given the high divergence between *SlX9* and *SlY9* (K_S = 14.4% and silent site divergence = 9.1%, see above), the *S. dioica* Y-linked sequence may be too diverged for the primers to work, or, alternatively, the Y copy may have been deleted in *S. dioica*.

All seven *S. dioica* sequences had the shorter haplotype (Figure 5-2). Compared with *S. latifolia* (see above), diversity appears to be slightly lower in *S. dioica* (from our sample of 7 X-linked sequences); using 538 silent sites, we have π = 1.6% (or 2.44% versus 2.95% in *S. latifolia*, for the 151 sites whose sequences could be compared in both species).

In terms of their sequences, the group of shorter *S. latifolia SlX9* haplotypes are more similar to the *S. dioica* sequence than the long ones. Divergence values of the short and long *SlX9* sequences from the *SdX9* sequences were 5.6% and 3.8%, respectively (net JC-corrected silent divergence values, based on 260 and 320 sites); net divergence values were 3.73% and 1.07%, respectively. Furthermore, the NJ tree (Figure 5-4) shows that the *S. latifolia* sequences with the short intron type cluster together with the *S. dioica* sequences, although all bootstrap values are very low. One *S. latifolia* individual (no. 32 in Table 5-2, from Northern France, which has the short intron 2 type), is very similar to a group of *SdX9* sequences (Figure 5-4), sharing four SNP variants with *S. dioica*, at sites at which all other *S. latifolia* sequences were fixed for a different variant. Introgression has thus probably occurred in its recent ancestry, which is plausible, as both species co-exist in this geographic region. Another *SlX9* sequence of the short intron type (from individual 6) is also similar to a group of *SdX9* sequences (Figure 5-4), but it comes from a region where *S. dioica* is not found (Southern France).

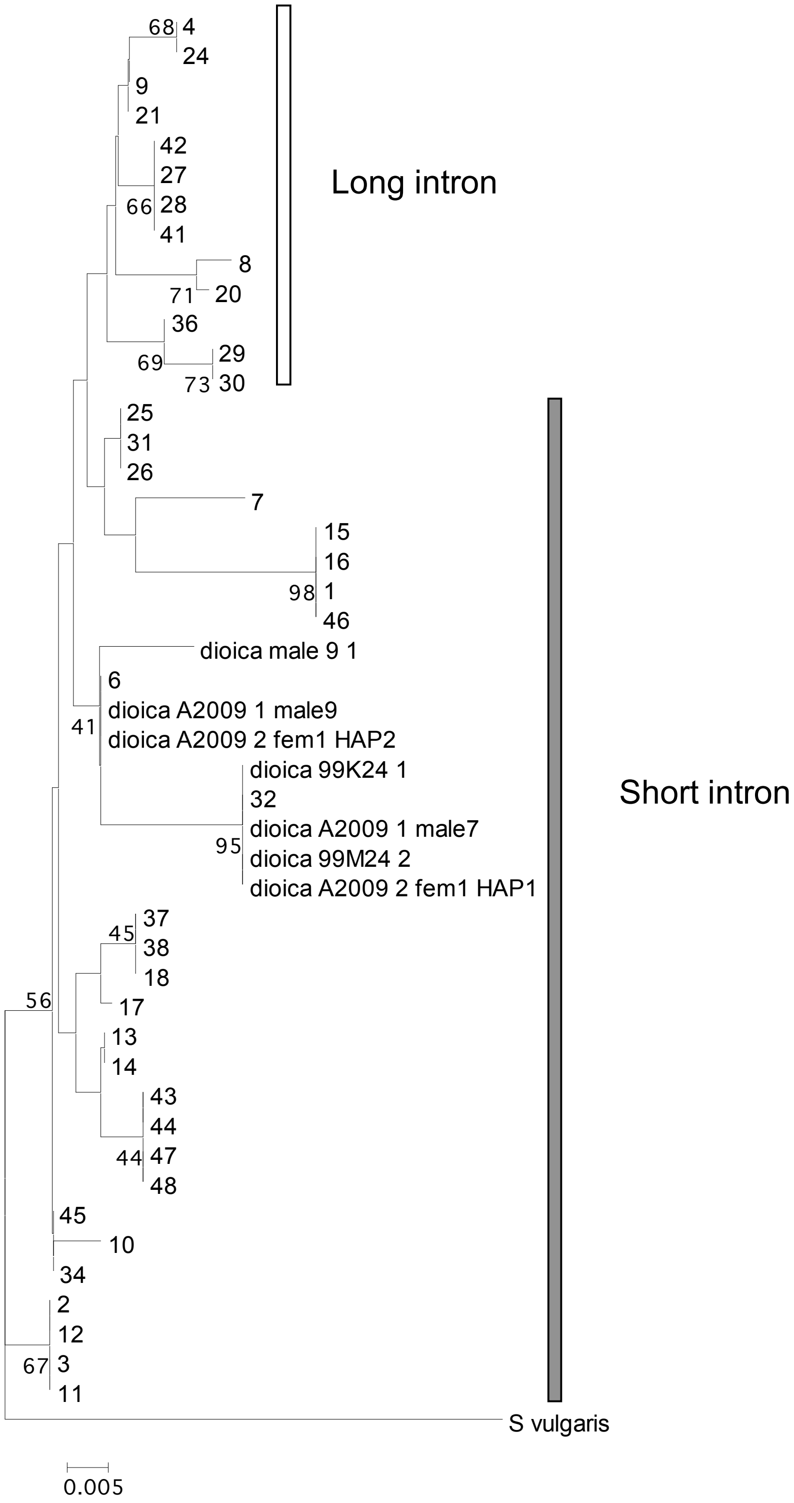


Figure 5-4: Neighbour-Joining tree constructed in MEGA, based on all sites, with bootstrap support values > 40% shown at the branches. Numbers represent *S. latifolia* individuals (see Table 5-2).

SlX9 and *SdX9* share 13 polymorphic sites, 7 of which are in regions shared among both *S. latifolia* haplotypes, while 5 are in the intron 2 region that is present only in the subset of the *S. latifolia* short haplotype group plus the *S. dioica* sequences (Supplementary Figure 5-1). There was one fixed differences between the sequences of the two species, in the very beginning of intron 1, but for this site, only four *S. latifolia* sequences aligned with *S. dioica* (Supplementary Figure 5-1, Table 5-7).

Table 5-7: The number of shared and fixed variants between *S. latifolia* and *S. dioica*. Note that, since we obtained 40 *S. latifolia* but only seven *S. dioica* sequences, the number of sites polymorphic in *S. latifolia* and fixed in *S. dioica* is probably an underestimate. This value, however, only includes sites for which we had all seven *S. dioica* sequences (i.e. excluding the last 558 bp of the alignment).

No. of fixed differences	No. of sites polymorphic in *S. latifolia*, but fixed in *S. dioica*	No. of sites polymorphic in *S. dioica*, but fixed in *S. latifolia*	No. of shared polymorphic sites
1	64	9	13

5.4.7 Linkage disequilibrium

Among 36 polymorphic sites in the *S. latifolia SlX9* gene, 9 pairs of sites are under significant linkage disequilibrium in our sample of alleles (Fisher's exact test; $p < 0.05$, after correction for multiple comparisons using the Bonferroni procedure). Eight of these pairs were less than 100bp apart from each other, whereas in one pair, the sites are 1493 bp apart (Table 5-8); all SNPs which were in significant LD, were downstream of exon 2, suggesting that there might be a small region of positive LD either within *SlX9* or extending downstream of the gene. The respective *S. dioica* sequences always corresponded to one of either SNP pair, i.e. the association between two nucleotides was also present in *S. dioica* (the *S. dioica* sample was polymorphic at one of these sites (pos. 2085), but we did not have all seven *S. dioica* haplotypes for all sites that were under LD in *S. latifolia* (see Supplementary Figure 5-1)).

Sites that were under significant LD in *S. latifolia* showed different patterns of polymorphism, depending on whether these sites were upstream or downstream of the intron 2 insertion (although this could be due to chance): all three sites upstream of this region were only polymorphic in the set of short *SlX9* sequences, whereas polymorphisms were shared across both intronic groups downstream of the insertion (Supplementary Figure 5-1). The three polymorphic sites upstream of the insertion (734, 794, 875) carried variants that were not found in our (limited) *S. dioica* sample. Note that polymorphisms that were *not* under significant LD were shared among the two *SlX9* haplotype groups also upstream of the length difference in intron 2.

The *ZnS* value, which measures the average linkage disequilibrium for all sites, was only 0.06, lower than the *ZnS* value for a different sex-linked gene, *SlX*1 (*ZnS* = 0.131) (ATANASSOV *et al.* 2001). The recombination parameter, *R*, between adjacent sites was 0.13, and the minimum number of recombination events (both estimated by DnaSP 4.0) was 8.

Table 5-8: Sites with significant linkage disequilibrium in the *SlX9*-diversity sample.

Position 1 (bp)	Position 2 (bp)	Distance (bp)
734	2227	1493
794	875	81
2016	2097	81
2016	2100	84
2067	2073	6
2085	2097	12
2085	2100	15
2097	2100	3
2154	2157	3

The alternative to introgression, ancestral polymorphism, with balancing selection in the common ancestor of the two species, and maintained in *S. latifolia*, predicts that we should observe many variants at higher frequencies than expected under neutrality, resulting in a positive Tajima's *D* statistic. However, in our sample, Tajima's *D* for *SlX9* was negative, though not significant (Tajima's *D* = -0.32, p > 0.1). Other tests of selection on *SlX9* were also not significant (Table 5-9), except for Fu's *F*, which is very sensitive to a frequency spectrum bias towards rare polymorphisms.

Table 5-9: Tests of neutrality and recombination for *SlX9* and *SlY9*.

Gene	Tajima's *D*	Fu and Li's *D**	Fu and Li's *F**	Fu's *F*	Strobeck's *S* statistic	Fay and Wu's *H*	No. of haplotypes	$R = 3N_er$ (adjacent sites)
SlX9	-0.32 NS	-0.25 NS	-0.32 NS	-7.63 $p < 0.05$	1.000	-1.23 NS	26 *p*(nH=28) < 0.001	0.14
SlY9	-1.19307 $p < 0.1$	- 1.74431 $p < 0.1$	- 1.84007 $p < 0.1$	- 0.515 NS	0.836	0.46377 NS	4 *p*(nHap=4) = 0.21	-

5.5 Discussion

5.5.1 Causes of low Y diversity, relative to the X

SlX9, is a sex-linked gene that shares characteristics with other known genes on the *S. latifolia* sex chromosomes. Although there is no overt sign that its Y-linked copy is degenerating, its diversity is reduced to a greater extent than expected based on the three times lower number of Y compared to X chromosomes in the population, similarly to the findings for all other Y-linked genes previously studied (*SlY4*, *DD44Y*, *SlY1* and *SlAp3Y*) (ATANASSOV *et al.* 2001; LAPORTE *et al.* 2005; MATSUNAGA *et al.* 2003). A reduced mutation rate on the Y cannot account for the diversity difference. Mutation rates in *S. latifolia* were found either not to differ between the X and Y copies (as we have found here), or to be significantly higher for the Y-linked copies (FILATOV 2005; FILATOV and CHARLESWORTH 2002; NICOLAS *et al.* 2005).

Population structure can increase species-wide diversity, but diversity on the Y chromosome should be more affected by subdivision than diversity on the X, even if pollen and seed dispersal rates are the same (LAPORTE and CHARLESWORTH 2002). Strong population structure has been found in *S. latifolia* for the genes *SlX4/Y4* and *DD44-X/Y* (LAPORTE *et al.* 2005)) and *SlX1/Y1* (ATANASSOV *et al.* 2001; IRONSIDE and FILATOV 2005), in all cases much more markedly for the Y than the X, due to the low Y diversity. Our small within-population samples are not suitable for rigorous tests for subdivision (only large-scale geographical subdivisions can be tested, see above), but our conclusion of reduced diversity on the Y should be conservative with respect to subdivision, and we can conclude that this probably cannot explain the high X diversity.

Accordingly, the data suggest that hitch-hiking processes have probably led to a reduced diversity of non-recombining regions of the *S. latifolia* Y chromosome. The Silene Y presumably contains many active genes, as suggested by its mere size (about 570 MB), and none of the eleven genes identified is degenerate, i.e. Hill-Robertson interference effects may be considerable. However, on a completely linked chromosome such as the Y, all sites are affected by the same selective or demographic processes, making it difficult to determine which might be the major force reducing Y diversity.

We observed an excess of low-frequency variants for *SlY9* (Table 5-9), as shown by a negative Tajima's *D*; however, with the small sample of genes whose diversity has been studied in this species, it is difficult to estimate and correct for any effects of recent demographic history that may have produced a tendency for such a frequency spectrum. (Note that the X-linked genes *SlssX*, *Sl-CypX* and *Sl-Cyt* also have negative Tajima's *D* values (BERGERO *et al.* 2008; Chapter 4; Bergero, unpublished data). The positive (though non-significant) Fay and Wu's *H* argues against a selective sweep on the Y chromosome. The data are, however, consistent with purifying selection and thus with results represented in Chapter 2 as well as the suggestion (BACHTROG 2008) that Muller's ratchet and/or background selection should be more important early in Y chromosome evolution, when the number of functional genes that can undergo detrimental mutations is still very large. Selective sweeps might become more important at intermediate stages of Y evolution, after the

Y has lost many genes, so that there is a higher chance that beneficial mutations can occur on chromosomes not carrying many deleterious mutations. Previous studies of three Y-linked genes (*SlY*1, *SlY*4 and *DD44Y*) did not find significant Tajima's *D* values, but argue that population subdivision could obscure the effects of a sweep (ATANASSOV *et al.* 2001; LAPORTE *et al.* 2005). On the other hand, the strong Y chromosome differentiation among populations (IRONSIDE and FILATOV 2005) argues against recent species-wide selective sweeps. Population structure is, however, consistent with background selection, or locally confined sweeps (IRONSIDE and FILATOV 2005).

5.5.2 Causes of high X diversity in *S. latifolia*

For *SlX9*, synonymous diversity is higher than for any other X-linked gene studied in this species, and introgression from *S. dioica* is certainly a plausible hypothesis to explain this. The shared short intron 2 structure, shared SNPs, and tendency for nearby sites to show positive linkage disequilibrium are consistent with introgression. Evidence for introgression from *S. dioica* into *S. latifolia* has also been reported for *DD44X* and *SlX4* (LAPORTE *et al.* 2005). For *SlX*4, a size variant in a *S. latifolia* intron matched an intron size variant in *S. dioica*, similar to our observation for *SlX9* (LAPORTE *et al.* 2005); for *SlX1*, nine out of ten shared polymorphisms were located in the first 1755 bp, and fixed differences were found only at the 3′ end of the gene (ATANASSOV *et al.* 2001), whereas, for *DD44X*, the sequences sampled from the two species shared 14 polymorphic sites and had no fixed differences (LAPORTE *et al.* 2005). These results suggest that the introgressed regions can be very localized (LAPORTE *et al.* 2005), and thus that it may be infrequent, and that the introgressed regions are often eliminated after recombination (presumably due to selection). Note that, for the diversity analysis of *SlX9*, exon 1 and intron 1 were not included, i.e. our diversity value might be an over-estimate if the region surveyed coincides with a region of the gene in which *SdX9* sequence has been introgressed.

Introgression between *S. latifolia* and *S. dioica* has also been detected using AFLP markers and (maternally inherited) chloroplast markers: out of 209 markers studied by MINDER *et al.* (2007), only 7 were species-specific, and five out of seven chloroplast haplotypes in *S. latifolia* were also present in *S. dioica* (PRENTICE *et al.*

2008). MINDER *et al.* (2007) did not observe significant LD between segregating AFLP markers, but these markers were scattered on different *S. latifolia* chromosomes, and were probably mostly loosely linked, so the chance of detecting LD would be slight unless a chromosome contains a large region that introgressed too recently to have recombined. Using a similar AFLP marker set, allopatric populations of *S. dioica* and *S. latifolia* in Switzerland separated by small distances were found to be more distinct than sympatric ones (MINDER and WIDMER 2008), as expected if hybridization occurs locally, but the introgressed regions are usually eliminated, rather than persisting.

For *SlX9*, the short sequences cluster together with *SdX9* in the phylogenetic tree, and form a distinct group from the long intronic sequences (Figure 5-4) (although the bootstrap support values are very low); in line with this, silent divergence from *S. dioica* is lower for the group of short *SlX9* sequences compared to the long ones, suggesting that introgression occurred recently enough to give these differences. Furthermore, individual no. 32, which comes from a region where *S. dioica* is present, carries a *SlX9* sequence that is very similar to those of *S. dioica* and might be derived from a recent introgression event.

However, the fact that we find both of the *SlX9* haplotypes in plants from the Mediterranean region, where *S. dioica* is absent, argues against a hypothesis of simple introgression as the source of the variant with the short intron 2. Since *S. latifolia* sequences of the long intron type also share variants at polymorphic sites with *S. dioica*, recombination must have occurred after introgression took place, to yield a group of haplotypes with the short intron 2 characteristic of *S. dioica*, but with variants derived from the long *S. latifolia* haplotype.

The finding of high diversity in *S. latifolia*, but not *S. dioica,* suggests introgression into *S. latifolia*, but not of *SlX9* sequences into *S. dioica* (SWEIGART and WILLIS 2003). This is consistent with the distribution of chloroplast versus genomic markers, which suggested that hybridization events mainly involve *S. dioica* as the pollen donor (MINDER *et al.* 2007). However, experiments with equal amounts of pollen from the two species yielded progeny from *S. latifolia* recipients in which less than 20% were hybrids, compared with 50% with *S. dioica* recipients (RAHME *et al.* 2009).

Since we do not have a sequence of the Y-linked copy of *SdX9* (if indeed a homologue exists in *S. dioica*), we cannot test for introgression of the Y-linked gene. The *SlY4* and *DD44-Y* genes show no signs of introgression, but instead, their sequences cluster by their species of origin (IRONSIDE and FILATOV 2005; LAPORTE *et al.* 2005). However, as LAPORTE *et al.* (2005) pointed out, the lower effective population size of the Y chromosome implies that shared polymorphisms are expected to be lost quickly, making introgression less detectable than for the X or the autosomes.

To estimate the diversity for X-linked genes relative to homologues on the Y (or to estimate X/autosome diversity ratio), it is clearly essential to have reliable diversity estimates for genes on the different chromosomes, and introgression from a different species will make this difficult and could increase the estimated diversity (SWEIGART and WILLIS 2003).

Is introgression the sole cause of higher X than Y diversity? If introgression of the X occurred recently, it might be possible to remove the introgressed alleles and estimate X diversity, for comparison with the Y. However, this may not be possible if introgression is not recent, so that one cannot recognise introgressed alleles. When we excluded from the *SlX9* dataset all polymorphisms that were shared with *S. dioica*, the X-Y difference in diversity remained unchanged, suggesting that introgression is not the sole factor for higher X than Y diversity (HKA test: χ^2 = 4.50, p <0.05). If most introgressed sequences are eliminated, and only certain small introgressed regions remain, this will be difficult to distinguish from shared ancestral polymorphisms for such closely related species. In the case of *Mimulus guttatus* and *M. nasutus*, introgression was detectable because high diversity was found only in those *M. guttatus* populations that are sympatric with *M. nasutus*, and also because *M. nasutus* sequences were quite readily distinguishable from *M. guttatus*, allowing recent introgression to be recognised (SWEIGART and WILLIS 2003). Both these characteristics differ from the situation for our species, and we cannot rule out that ancestral polymorphism has contributed to some of the observed pattern in *S. latifolia*. It does seem unlikely, however, that ancestral polymorphism is detectable in *S. latifolia* only, and not in *S. dioica*.

5.5.3 Introgression and X-Y divergence

If *SlX9* and *SlY9* stopped recombining after *S. latifolia* and *S. dioica* split into two species, the X-linked copy of one species should be more similar to that species' Y-linked copy than to the X of the other. Introgression could then inflate the divergence between X- and Y-linked copies, relative to the actual time since they stopped recombining, and different degrees of introgression of different regions of the X could make it difficult to determine the true times when recombination stopped, at least for regions in which it stopped most recently.

However, this is unlikely to have affected our results for *SlXY9*, since we can say for *Sl-Cyp*, the gene most closely linked to *SlX9/Y9*, that it probably stopped recombining before the two species split: *Sl-CypY* carries an intronic MITE insterion in *S. latifolia* and *S. dioica*, but the insertion is absent from *Sl-CypX*. Furthermore, *SlX9-SlY9* divergence is much higher than that between the two species, making a scenario of independent recombination cessation unlikely.

6 Conclusion and future directions

The results presented in this thesis show that purifying selection acting against deleterious mutations can cause patterns of diversity and rates of degeneration that are consistent with data from Drosophila, suggesting that levels of diversity on the *D. melanogaster* fourth (dot) chromosome and the *D. miranda* neo-Y chromosome can be explained by selection acting against deleterious mutations only. Similarly, levels of diversity on the *D. americana* dot chromosome are only about 10-17 fold reduced compared to the autosomes (BETANCOURT *et al.* 2009), confirming the relatively high diversity in regions of low recombination in Drosophila. These results are in contrast to previous findings which suggested that the strong skew in the frequency spectrum at segregating sites observed for the neo-Y, in combination with its reduced diversity, is not compatible with background selection (BACHTROG 2004). This conclusion, however, was based on the assumption that background selection reduces diversity in a deterministic fashion, and the work presented in chapter 2 shows that this is not the case. If interference effects are taken into account, the observed distortions in the frequency spectra at segregating sites are indeed compatible with purifying selection alone, and hence with the Drosophila data. The new results also suggest that, with an increasing number of sites linked on a non-recombining chromosome, purifying selection against relatively strongly deleterious mutations becomes increasingly less efficient, resulting in a situation where the effective population size (and hence neutral diversity) is larger than predicted by the current model of background selection, which assumes independence among sites (NORDBORG *et al.* 1996). Mathematical models taking these interference effects into account are yet to be developed.

The validity of Muller's ratchet in driving the degeneration of asexual population has previously been questioned, mainly because the time-scales involved were estimated to be too large. In chapter 3, it was shown that, based on estimates of mutation rates and selection coefficients against loss-of-function mutations, the predicted rate of accumulation of such mutations is consistent with the rate observed for the evolving neo-Y chromosome of *D. miranda.* Furthermore, selection at "nonsynonymous" sites can accelerate the process of gene loss, an effect that had not been investigated before.

The model presented could be extended by allowing weak selection at "synonymous" sites or by modelling the effects of strongly beneficial mutations. Even if the data from Drosophila are compatible with purifying selection only, we can realistically assume that some adaptive changes do occur on evolving Y chromosomes. For example, the human and Drosophila Y have accumulated male-function genes (CARVALHO *et al.* 2009; SKALETSKY *et al.* 2003), and this most likely happened most by positive selection, though little is known about the time-scales involved (i.e. if these genes accumulated after other genes had been lost already). Adding positive selection to the model is likely to speed up degeneration, though it is questionable if neutral diversity would be reduced much further; this is because beneficial mutations are, relatively speaking, just the opposite of deleterious mutations, and the plateauing effect of selection on N_e might persist – the exact outcome would most likely depend considerably on the selection coefficients assumed.

In chapters 4 and 5, I have described the discovery of two new sex-linked genes in *S. latifolia*, *SlCyt* and *SlX9/SlY9,* both of which were mapped onto the sex chromosomes. I have shown that *SlCyt* moved onto the X only after the sex chromosomes had evolved, possibly after the split between *S. latifolia/S. dioica* and *S. diclinis*. *SlCyt* is the first known gene to have moved onto the Silene X chromosome and is now situated in a region that stopped recombining with the Y only recently. This, together with evidence of a selective sweep affecting diversity at *SlCyt*, raises the question of whether the translocation itself might have caused recombination suppression in the genomic region; more data are needed to resolve this question. *SlX9/Y9* shows similar characteristics to other known genes on the Silene sex chromosomes, such as reduced Y diversity and possibly introgression from *S. diocia*, which might have inflated diversity.

The theoretical results presented in the first chapters suggest that purifying selection can explain the Drosophila data - but how do they relate to Y chromosome evolution in *S. latifolia*? Given the molecular data that are available, the *S. latifolia* Y chromosome seems less degenerated than the *D. miranda* neo-Y: except for *Sl-Cyt* (which has been transposed from an autosome, see chapter 4), and *SlssY* (which might contain nonsynonymous substitutions that impair protein function (FILATOV 2008)), all other X-linked genes now described have intact Y-linked homolgues, and all Y-linked genes that have been investigated are expressed in males. However, some of the previous studies may have been biased towards finding intact genes on the Y, by screening cDNA from male tissue, using Y-derived probes (ATANASSOV *et al.* 2001; DELICHERE *et al.* 1999; MOORE *et al.* 2003); hence, only genes that were actually present on the Y could be detected. We cannot exclude the possibility that the Silene Y has lost many of its genes, similar to the situation on the *D. miranda* neo-Y - the fact that Silene plants without an X chromosome are not viable, suggests that at least some degeneration has taken place. The approach used here, i.e. searching for X-linked genes and scoring their segregation pattern in a mapping family, might be a better means to assess the amount of Y degeneration. Indeed, this is the first time that an X-linked gene has been found which lacks a Y-linked homologue. So far, however, the sample size of X-Y pairs available is too small to deduce any general pattern of the amount of degeneration; new sequencing technology is likely to help to search for genes more time-efficiently.

More information on parameters of mutation and selection need to be known in order to assess the evolutionary forces that shape the Silene Y. For example, to investigate the speed of degeneration, better estimates of the age of the system are needed; we do not have any estimates of the effective population size in Silene, nor its (deleterious) mutation rate and the distribution of mutational effects. We also need to understand the biology of the system better: For example, an important difference between the *D. miranda* neo-Y and the Silene Y, which is likely to affect degeneration, might be the amount of Y-linked genes expressed in the haploid stage. Since the Silene Y is derived from an ordinary autosome, the proportion of genes expressed in pollen is likely to high; in Arabidopsis, 61% of genes expressed in the sporophyte were also detected as mRNAs in pollen (HONYS and TWELL 2003). This could effectively increase selection against deleterious mutations, and slow down degeneration. As is known from Drosophila, sheltering of Y-linked mutations when there is no haploid expression can reduce selection against deleterious mutations considerably (CROW and SIMMONS 1983). This might also explain the higher diversity on the Silene Y chromosome compared to the *D. miranda* neo-Y because stronger selection leads to higher diversity under BGS.

Due to gene expression in pollen, the relative importance of positive versus negative selection might also differ between plants and animals. Assuming that at least some beneficial mutations are recessive, positive selection might be more efficient on plant Y chromosomes compared to animals. Similarly, nothing is known about dosage compensation in Silene or in any other plant sex chromosome system; dosage compensation might lead to faster degeneration, again, because deleterious mutations on the Y are sheltered more efficiently. In Silene, the sex chromosomes evolved *de novo* (so a dosage compensation mechanism was probably not in place when the sex chromosomes evolved), whereas in *D. miranda* (which already contained a Y chromosome), the dosage compensation machinery might have been recruited to new chromosomal regions more readily. To establish whether there is dosage compensation in Silene, it is, of course, first it is necessary to find Y-degenerate genes before expression analyses can be done.

Hence, there is still a lot to be learnt from the Silene system. It might turn out that, even though many aspects of sex chromosome evolution are surprisingly similar among different species (such as a step-wise recombination cessation between the X and Y or reduced Y diversity), other aspects, such as the speed of degeneration or the processes driving degeneration might differ, depending on the biology of the system in question.

7 Bibliography

1 Introduction

BACHTROG, D., E. HOM, K. M. WONG, X. MASIDE and P. DE JONG, 2008 Genomic degradation of a young Y chromosome in Drosophila miranda. Genome Biology **9:** R30.

BARRACLOUGH, T. G., D. FONTANETO, C. RICCI and E. A. HERNIOU, 2007 Evidence for inefficient selection against deleterious mutations in cytochrome oxidase I of asexual bdelloid rotifers. Molecular Biology and Evolution **24:** 1952-1962.

BARTOLOME, C., and B. CHARLESWORTH, 2006 Evolution of amino-acid sequences and codon usage on the Drosophila miranda neo-sex chromosomes. Genetics **174:** 2033-2044.

BARTON, N. H., and B. CHARLESWORTH, 1998 Why sex and recombination? Science **281:** 1986-1990.

BARTON, N. H., and S. P. OTTO, 2005 Evolution of recombination due to random drift. Genetics **169:** 2353-2370.

BERGERO, R., and D. CHARLESWORTH, 2009 The evolution of restricted recombination in sex chromosomes. Trends Ecol Evol **24:** 94-102.

BERGERO, R., D. CHARLESWORTH, D. A. FILATOV and R. C. MOORE, 2008a Defining regions and rearrangements of the Silene latifolia Y chromosome. Genetics **178:** 2045-2053.

BERGERO, R., A. FORREST and D. CHARLESWORTH, 2008b Active miniature transposons from a plant genome and its nonrecombining Y chromosome. Genetics **178:** 1085-1092.

BERGERO, R., A. FORREST, E. KAMAU and D. CHARLESWORTH, 2007 Evolutionary Strata on the X Chromosomes of the Dioecious Plant Silene latifolia: Evidence From New Sex-Linked Genes. Genetics **175:** 1945-1954.

BETANCOURT, A. J., J. J. WELCH and B. CHARLESWORTH, 2009 Reduced Effectiveness of Selection Caused by a Lack of Recombination. Current Biology **19:** 655-660.

BROSSEAU, G. E., 1960 Genetic Analysis of the Male Fertility Factors on the Y-Chromosome of Drosophila-Melanogaster. Genetics **45:** 257-274.

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

CARVALHO, A. B., L. B. KOERICH and A. G. CLARK, 2009 Origin and evolution of Y chromosomes: Drosophila tales. Trends in Genetics **25:** 270-277.

CHARLESWORTH, B., 1994 The Effect of Background Selection against Deleterious Mutations on Weakly Selected, Linked Variants. Genetical Research **63:** 213-227.

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

CHARLESWORTH, B., 2009 Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics **10:** 195-205.

CHARLESWORTH, B., and D. CHARLESWORTH, 1978 Model for 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., 2002 Plant sex determination and sex chromosomes. Heredity **88:** 94-101.

CHARLESWORTH, D., and B. CHARLESWORTH, 2005 Sex chromosomes: Evolution of the weird and wonderful. Current Biology **15:** R129-R131.

CROW, J. F., and M. KIMURA, 1970 An Introduction to Population Genetics Theory. An Introduction to Population Genetics Theory**:** 591.

ERLANDSSON, R., J. WILSON and S. PÄÄBO, 2000 Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans. Mol Biol Evol **17:** 804-812.

EWENS, W. J., 2004 *Mathematical Population Genetics*. Springer-Verlag, Berlin.

EYRE-WALKER, A., and P. D. KEIGHTLEY, 1999 High genomic deleterious mutation rates in hominids. Nature **397:** 344-347.

EYRE-WALKER, A., and P. D. KEIGHTLEY, 2007 The distribution of fitness effects of new mutations. Nature Reviews Genetics **8:** 610-618.

EYRE-WALKER, A., M. WOOLFIT and T. PHELPS, 2006 The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics **173:** 891-900.

FELSENSTEIN, J., 1974 Evolutionary Advantage of Recombination. Genetics **78:** 737-756.

FILATOV, D. A., 2005 Evolutionary history of Silene latifola sex chromosomes revealed by genetic mapping of four genes. Genetics **170:** 975-979.

FILATOV, D. A., 2008 A selective sweep in or near the *Silene latifolia* X-linked gene SlssX. Genetical Research **90:** 85-95.

FISHER, R. A., 1930 *The genetical theory of natural selection.* New York: Dover Publications, 1958.

FRASER, J. A., S. DIEZMANN, R. L. SUBARAN, A. ALLEN, K. B. LENGELER *et al.*, 2004 Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. Plos Biology **2:** 2243-2255.

GETHMANN, R. C., 1988 Crossing over in Males of Higher Diptera (Brachycera). Journal of Heredity **79:** 344-350.

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.

HAAG-LIAUTARD, C., M. DORRIS, X. MASIDE, S. MACASKILL, D. L. HALLIGAN *et al.*, 2007 Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. Nature **445:** 82-85.

HADANY, L., and M. W. FELDMAN, 2005 Evolutionary traction: the cost of adaptation and the evolution of sex. Journal of Evolutionary Biology **18:** 309-314.

HADDRILL, P. R., D. L. HALLIGAN, D. TOMARAS and B. CHARLESWORTH, 2007 Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. Genome Biology **8**.

HAIGH, J., 1978 Accumulation of Deleterious Genes in a Population - Mullers Ratchet. Theoretical Population Biology **14:** 251-267.

HELLMANN, I., K. PRUFER, H. K. JI, M. C. ZODY, S. PAABO *et al.*, 2005 Why do human diversity levels vary at a megabase scale? Genome Research **15:** 1222-1231.

HILL, W. G., and A. ROBERTSON, 1966 Effect of Linkage on Limits to Artificial Selection. Genetical Research **8:** 269-294.

JENSEN, M. A., B. CHARLESWORTH and M. KREITMAN, 2002 Patterns of genetic variation at a chromosome 4 locus of Drosophila melanogaster and D-simulans. Genetics **160:** 493-507.

JOHNSON, T., and N. H. BARTON, 2002 The effect of deleterious alleles on adaptation in asexual populations. Genetics **162:** 395-411.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The Hitchhiking Effect Revisited. Genetics **123:** 887-899.

KEIGHTLEY, P. D., and S. P. OTTO, 2006 Interference among deleterious mutations favours sex and recombination in finite populations. Nature **443:** 89-92.

KIMURA, M., 1983 The neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

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

LAWSON-HANDLEY, L., H. CEPLITIS and H. ELLEGREN, 2004 Evolutionary strata on the chicken Z chromosome: Implications for sex chromosome evolution. Genetics **167:** 367-376.

LIU, Z. Y., P. H. MOORE, H. MA, C. M. ACKERMAN, M. RAGIBA *et al.*, 2004 A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature **427:** 348-352.

LOEWE, L., and B. CHARLESWORTH, 2006 Inferring the distribution of mutational effects on fitness in Drosophila. Biology Letters **2:** 426-430.

LOEWE, L., B. CHARLESWORTH, C. BARTOLOME and V. NOEL, 2006 Estimating selection on nonsynonymous mutations. Genetics **172:** 1079-1092.

MARAIS, G. A. B., M. NICOLAS, R. BERGERO, P. CHAMBRIER, E. KEJNOVSKY *et al.*, 2008 Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. Current Biology **18:** 545-549.

MATSUBARA, K., H. TARUI, M. TORIBA, K. YAMADA, C. NISHIDA-UMEHARA *et al.*, 2006 Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. Proceedings of the National Academy of Sciences of the United States of America **103:** 18190-18195.

MATSUNAGA, S., 2006 Sex chromosome-linked genes in plants. Genes & Genetic Systems **81:** 219-226.

MATSUNAGA, S., E. ISONO, E. KEJNOVSKY, B. VYSKOT, J. DOLEZEL *et al.*, 2003 Duplicative transfer of a MADS box gene to a plant Y chromosome. Molecular Biology and Evolution **20:** 1062-1069.

MAYNARD SMITH, J., 1978 *The Evolution of Sex*. Cambridge University Press.

MAYNARD SMITH, J., and J. HAIGH, 1974 Hitch-Hiking Effect of a Favorable Gene. Genetical Research **23:** 23-35.

MCVEAN, G. A. T., 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.

MORGAN, T. H., 1913 *Heredity and Sex*. Columbia University Press, New York.

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

MULLER, H. J., 1964 The Relation of Recombination to Mutational Advance. Mutation Research **1:** 2-9.

MYERS, S., L. BOTTOLO, C. FREEMAN, G. A. T. MCVEAN and P. DONNELLY, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321-324.

NAM, K., and H. ELLEGREN, 2008 The Chicken (*Gallus gallus*) Z Chromosome Contains at Least Three Nonlinear Evolutionary Strata. Genetics **180:** 1131-1136.

NEI, M., 1969 Linkage modification and sex difference in recombination. Genetics **63:** 681-699.

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.

OTTO, S. P., and N. H. BARTON, 1997 The evolution of recombination: Removing the limits to natural selection. Genetics **147:** 879-906.

PALAND, S., and M. LYNCH, 2006 Transitions to asexuality result in excess amino acid substitutions. Science **311:** 990-992.

PEICHEL, C. L., J. A. ROSS, C. K. MATSON, M. DICKSON, J. GRIMWOOD *et al.*, 2004 The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. Current Biology **14:** 1416-1424.

RICE, W. R., 1984 Sex-chromosomes and the evolution of sexual dimorphism. Evolution **38:** 735-742.

ROZE, D., and N. H. BARTON, 2006 The Hill-Robertson effect and the evolution of recombination. Genetics **173:** 1793-1811.

SANCHEZ, L., 2008 Sex-determining mechanisms in insects. International Journal of Developmental Biology **52:** 837-856.

SHETTY, S., D. K. GRIFFIN and J. A. M. GRAVES, 1999 Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. Chromosome Research **7:** 289-295.

SKALETSKY, H., T. KURODA-KAWAGUCHI, P. J. MINX, H. S. CORDUM, L. HILLIER *et al.*, 2003 The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423:** 825-U822.

STEINEMANN, M., and S. STEINEMANN, 1998 Enigma of Y chromosome degeneration: Neo-Y and Neo-X chromosomes of Drosophila miranda a model for sex chromosome evolution. Genetica **102-3:** 409-420.

STEINEMANN, S., and M. STEINEMANN, 2005 Y chromosomes: born to be destroyed. Bioessays **27:** 1076-1083.

VAN TUINEN, M., and S. B. HEDGES, 2001 Calibration of avian molecular clocks. Molecular Biology and Evolution **18:** 206-213.

VEUSKENS, J., D. YE, M. OLIVEIRA, D. D. CIUPERCESCU, P. INSTALLE *et al.*, 1992 Sex Determination in the dioecious *Melandrium album* - androgenic embryogenesis requires the presence of the X-chromosome. Genome **35:** 8-16.

WANG, W., K. THORNTON, A. BERRY and M. Y. LONG, 2002 Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science **295:** 134-137.

WEST, S. A., C. M. LIVELY and A. F. READ, 1999 A pluralist approach to sex and recombination. Journal of Evolutionary Biology **12:** 1003-1012.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 0097-0159.

WRIGHT, S. I., J. P. FOXE, L. DEROSE-WILSON, A. KAWABE, M. LOOSELEY *et al.*, 2006 Testing for effects of recombination rate on nucleotide diversity in natural populations of Arabidopsis lyrata. Genetics **174:** 1421-1430.

ANDOLFATTO, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in Drosophila melanogaster and Drosophila simulans. Molecular Biology and Evolution **18:** 279-290.

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

BACHTROG, D., E. HOM, K. M. WONG, X. MASIDE and P. DE JONG, 2008 Genomic degradation of a young Y chromosome in Drosophila miranda. Genome Biology **9:** R30.

BARTOLOMÉ, C., and B. CHARLESWORTH, 2006 Evolution of amino-acid sequences and codon usage on the *Drosophila miranda* neo-sex chromosomes. Genetics **174:** 2033-2044.

BIRKY, C. W., and J. B. WALSH, 1988 Effects of Linkage on Rates of Molecular Evolution. Proceedings of the National Academy of Sciences of the United States of America **85:** 6414-6418.

CHARLESWORTH, B., 1996 Background selection and patterns of genetic diversity in Drosophila melanogaster. Genetical Research **68:** 131-149.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993a The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics **134:** 1289-1303.

CHARLESWORTH, D., M. T. MORGAN and B. CHARLESWORTH, 1992 The effect of linkage and population size on inbreeding depression due to mutational load. Genet. Res. **59:** 49-61.

CHARLESWORTH, D., M. T. MORGAN and B. CHARLESWORTH, 1993b Mutation accumulation in finite outbreeding and inbreeding populations. Genet. Res. **61:** 39-56.

COMERON, J. M., A. WILLIFORD and R. M. KLIMAN, 2008 The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity **100:** 19-31.

FELSENSTEIN, J., 1974 Evolutionary Advantage of Recombination. Genetics **78:** 737-756.

HALDANE, J. B. S., 1927 A mathematical theory of natural and artificial selection, Part V: Selection and mutation. Proceedings of the Cambridge Philosophical Society **23:** 838-844.

HILL, W. G., and A. ROBERTSON, 1966 Effect of Linkage on Limits to Artificial Selection. Genetical Research **8:** 269-294.

HILLIKER, A. J., G. HARAUZ, A. G. REAUME, M. GRAY, S. H. CLARK *et al.*, 1994 Meiotic Gene Conversion Tract Length Distribution within the Rosy Locus of Drosophila-Melanogaster. Genetics **137:** 1019-1024.

HUDSON, R. R., and N. L. KAPLAN, 1995 Deleterious background selection with recombination. Genetics **141:** 1605-1617.

KEIGHTLEY, P. D., and A. EYRE-WALKER, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics**:** 2251-2261.

KIMURA, M., 1983 The neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

LOEWE, L., and B. CHARLESWORTH, 2006 Inferring the distribution of mutational effects on fitness in Drosophila. Biol. Lett. **2:** 426-430.

LOEWE, L., and B. CHARLESWORTH, 2007 Background selection in single genes may explain patterns of codon bias. Genetics **175:** 1381-1393.

MARUYAMA, T., and M. KIMURA, 1974 A Note on the Speed of Gene Frequency Changes in Reverse Directions in a Finite Population. Evolution **28:** 161-163.

MCVEAN, G. A. T., and B. CHARLESWORTH, 1999 A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genetical Research **74:** 145-158.

MCVEAN, G. A. T., 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.

NORDBORG, M., B. CHARLESWORTH and D. CHARLESWORTH, 1996 The effect of recombination on background selection. Genetical Research **67:** 159-174.

ORR, H. A., 2000 The rate of adaptation in asexuals. Genetics **155:** 961-968.

PALSSON, S., 2004 On the effects of background selection in small populations on comparisons of molecular variation. Hereditas **141:** 74-80.

SCHAEFFER, S. W., 2002 Molecular population genetics of sequence length diversity in the Adh region of Drosophila pseudoobscura. Genetical Research **80:** 163-175.

SHAPIRO, J. A., W. HUANG, C. ZHANG, M. HUBISZ, J. LU *et al.*, 2007 Adaptive genic evolution in the Drosophila genome. Proc. Natl. Acad. Sci. USA **104:** 2271-2276.

SHELDAHL, L. A., D. M. WEINREICH and D. M. RAND, 2003 Recombination, dominance and selection on amino acid polymorphism, in the Drosophila genome: Contrasting patterns on the X and fourth chromosomes. Genetics **165:** 1195-1208.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

WANG, W., K. THORNTON, A. BERRY and M. Y. LONG, 2002 Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science **295:** 134-137.

3 The rate of gene loss on the *Drosophila miranda* neo-Y chromosome can be explained by the process of Muller's ratchet

ANDOLFATTO, P., 2007 Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genome Research **17:** 1755-1762.

BACHTROG, D., 2003 Protein evolution and codon usage bias on the neo-sex chromosomes of Drosophila miranda. Genetics **165:** 1221-1232.

BACHTROG, D., 2005 Sex chromosome evolution: Molecular aspects of Y-chromosome degeneration in Drosophila. Genome Research **15:** 1393-1401.

BACHTROG, D., 2006 Expression profile of a degenerating neo-Y chromosome in Drosophila. Current Biology **16:** 1694-1699.

BACHTROG, D., 2007 Reduced selection for codon usage bias in Drosophila miranda. Journal of Molecular Evolution **64:** 586-590.

BACHTROG, D., 2008a Similar rates of protein adaptation in Drosophila miranda and D-melanogaster, two species with different current effective population sizes BMC EVOLUTIONARY BIOLOGY **8:** article no. 334.

BACHTROG, D., 2008b The temporal dynamics of processes underlying Y chromosome degeneration. Genetics **179:** 1513-1525.

BACHTROG, D., and P. ANDOLFATTO, 2006 Selection, recombination and demographic history in Drosophila miranda. Genetics **174:** 2045-2059.

BACHTROG, D., E. HOM, K. M. WONG, X. MASIDE and P. DE JONG, 2008 Genomic degradation of a young Y chromosome in Drosophila miranda. Genome Biology **9:** R30.

BARTOLOMÉ, C., and B. CHARLESWORTH, 2006 Evolution of amino-acid sequences and codon usage on the *Drosophila miranda* neo-sex chromosomes. Genetics **174:** 2033-2044.

BELL, G., 1988 Recombination and the Immortality of the Germ Line. Journal of Evolutionary Biology **1:** 67-82.

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.

BETANCOURT, A. J., J. J. WELCH and B. CHARLESWORTH, 2009 Reduced Effectiveness of Selection Caused by a Lack of Recombination. Current Biology **19:** 655-660.

BONE, J. R., and M. I. KURODA, 1996 Dosage compensation regulatory proteins and the evolution of sex chromosomes in drosophila. Genetics **144:** 705-713.

CARVALHO, A. B., L. B. KOERICH and A. G. CLARK, 2009 Origin and evolution of Y chromosomes: Drosophila tales. Trends in Genetics **25:** 270-277.

CHARLESWORTH, B., 1978 Model for Evolution of Y-Chromosomes and Dosage Compensation. Proceedings of the National Academy of Sciences of the United States of America **75:** 5618-5622.

CHARLESWORTH, B., 1994 The Effect of Background Selection against Deleterious Mutations on Weakly Selected, Linked Variants. Genetical Research **63:** 213-227.

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

CHARLESWORTH, B., 2009 Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics **10:** 195-205.

CHARLESWORTH, B., and D. CHARLESWORTH, 1978 Model for Evolution of Dioecy and Gynodioecy. American Naturalist **112:** 975-997.

CHARLESWORTH, B., and D. CHARLESWORTH, 1997 Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genetical Research **70:** 63-73.

CHARLESWORTH, B., and D. CHARLESWORTH, 1998 Some evolutionary consequences of deleterious mutations. Genetica **103:** 3-19.

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., and K. A. HUGHES, 1999 The Maintenance of Genetic Variation in Life-History Traits, pp. pp. 369-392 in *Evolutionary Genetics: From Molecules to Morphology*, edited by R. S. SINGH and C. B. KRIMBAS. Cambridge University Press.

CLARK, A. G., M. B. EISEN, D. R. SMITH, C. M. BERGMAN, B. OLIVER *et al.*, 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature **450:** 203-218.

CROW, J. F., and M. J. SIMMONS, 1983 *The mutation load in Drosophila. In: The genetics and biology of Drosophila*

. Academic Press, London.

ENGELSTÄDTER, J., 2008 Muller's Ratchet and the Degeneration of Y Chromosomes: A Simulation Study. Genetics **180:** 957-967.

EWENS, W. J., 2004 *Mathematical Population Genetics.* Springer-Verlag, Berlin.

FELSENSTEIN, J., 1974 Evolutionary Advantage of Recombination. Genetics **78:** 737-756.

FISHER, R. A., 1930 *The genetical theory of natural selection.* New York : Dover Publications, 1958.

FRIDOLFSSON, A. K., H. CHENG, N. G. COPELAND, N. A. JENKINS, H. C. LIU *et al.*, 1998 Evolution of the avian sex chromosomes from an ancestral pair of autosomes. Proceedings of the National Academy of Sciences of the United States of America **95:** 8147-8152.

GESSLER, D. D. G., 1995 The constraints of finite size in asexual populations and the rate of the ratchet. Genetical Research **66:** 241-253.

GETHMANN, R. C., 1988 Crossing over in Males of Higher Diptera (Brachycera). Journal of Heredity **79:** 344-350.

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

GORDO, I., and B. CHARLESWORTH, 2000b The degeneration of asexual haploid populations and the speed of Muller's ratchet. Genetics **154:** 1379-1387.

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., A. NAVARRO and B. CHARLESWORTH, 2002 Muller's ratchet and the pattern of variation at a neutral locus. Genetics **161:** 835-848.

HAAG-LIAUTARD, C., M. DORRIS, X. MASIDE, S. MACASKILL, D. L. HALLIGAN *et al.*, 2007 Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. Nature **445:** 82-85.

HADANY, L., and M. W. FELDMAN, 2005 Evolutionary traction: the cost of adaptation and the evolution of sex. Journal of Evolutionary Biology **18:** 309-314.

HAIGH, J., 1978 Accumulation of Deleterious Genes in a Population - Mullers Ratchet. Theoretical Population Biology **14:** 251-267.

HALDANE, J. B. S., 1927 A mathematical theory of natural and artificial selection, Part V: Selection and mutation. Proceedings of the Cambridge Philosophical Society **23:** 838-844.

HIGGS, P. G., and G. WOODCOCK, 1995 The Accumulation of Mutations in Asexual Populations and the Structure of Genealogical Trees in the Presence of Selection. Journal of Mathematical Biology **33:** 677-702.

HILL, W. G., and A. ROBERTSON, 1966 Effect of Linkage on Limits to Artificial Selection. Genetical Research **8:** 269-294.

HUDSON, R. R., 1990 Gene Genealogies and the Coalescent Process. Futuyma, D. and J. Antonovics (Ed.). Oxford Surveys in Evolutionary Biology, Vol. 7. Xiv+314p. Oxford University Press: Oxford, England, Uk; New York, New York, USA. Illus. Maps**:** 1-44.

JAIN, K., 2008 Loss of least-loaded class in asexual populations due to drift and epistasis. Genetics **179:** 2125-2134.

JOHNSON, T., and N. H. BARTON, 2002 The effect of deleterious alleles on adaptation in asexual populations. Genetics **162:** 395-411.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The Hitchhiking Effect Revisited. Genetics **123:** 887-899.

KIMURA, M., 1962 On Probability of Fixation of Mutant Genes in a Population. Genetics **47:** 713-&.

KIMURA, M., 1983 The neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

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

LANGLEY, C. H., R. A. VOELKER, A. J. L. BROWN, S. OHNISHI, B. DICKSON *et al.*, 1981 Null Allele Frequencies at Allozyme Loci in Natural-Populations of Drosophila-Melanogaster. Genetics **99:** 151-156.

LOEWE, L., and B. CHARLESWORTH, 2007 Background selection in single genes may explain patterns of codon bias. Genetics **175:** 1381-1393.

LOEWE, L., B. CHARLESWORTH, C. BARTOLOME and V. NOEL, 2006 Estimating selection on nonsynonymous mutations. Genetics **172:** 1079-1092.

MARIN, I., A. FRANKE, G. J. BASHAW and B. S. BAKER, 1996 The dosage compensation system of Drosophila is co-opted by newly evolved X chromosomes. Nature **383:** 160-163.

MAYNARD SMITH, J., and J. HAIGH, 1974 Hitch-Hiking Effect of a Favorable Gene. Genetical Research **23:** 23-35.

MCVEAN, G. A. T., 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.

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

MULLER, H. J., 1964 The Relation of Recombination to Mutational Advance. Mutation Research **1:** 2-9.

PAMILO, P., M. NEI and W. H. LI, 1987 Accumulation of Mutations in Sexual and Asexual Populations. Genetical Research **49:** 135-146.

RICE, W. R., 1987 Genetic Hitchhiking and the Evolution of Reduced Genetic-Activity of the Y-Sex Chromosome. Genetics **116:** 161-167.

ROUZINE, I. M., E. BRUNET and C. O. WILKE, 2008 The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. Theoretical Population Biology **73:** 24-46.

SAXENA, R., L. G. BROWN, T. HAWKINS, R. K. ALAGAPPAN, H. SKALETSKY *et al.*, 1996 The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. Nature Genetics **14:** 292-299.

SELLA, G., D. A. PETROV, M. PRZEWORSKI and P. ANDOLFATTO, 2009 Pervasive Natural Selection in the Drosophila Genome? Plos Genetics **5**.

SKALETSKY, H., T. KURODA-KAWAGUCHI, P. J. MINX, H. S. CORDUM, L. HILLIER *et al.*, 2003 The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423:** 825-U822.

SÖDERBERG, R. J., and O. G. BERG, 2007 Mutational interference and the progression of Muller's ratchet when mutations have a broad range of deleterious effects. Genetics **177:** 971-986.

STEINEMANN, M., and S. STEINEMANN, 1998 Enigma of Y chromosome degeneration: Neo-Y and Neo-X chromosomes of Drosophila miranda a model for sex chromosome evolution. Genetica **102-3:** 409-420.

STEINEMANN, S., and M. STEINEMANN, 1999 The Amylase gene cluster on the evolving sex chromosomes of Drosophila miranda. Genetics **151:** 151-161.

STEPHAN, W., L. CHAO and J. G. SMALE, 1993 The Advance of Muller Ratchet in a Haploid Asexual Population - Approximate Solutions Based on Diffusion-Theory. Genetical Research **61:** 225-231.

YI, S. J., D. BACHTROG and B. CHARLESWORTH, 2003 A survey of chromosomal and nucleotide sequence variation in Drosophila miranda. Genetics **164:** 1369-1381.

4 *Slcyt*, a newly identified sex-linked gene, has recently moved onto the X chromosome in *Silene latifolia* (Caryophyllaceae)

ASHBURNER, M., K. G. GOLIC and R. S. HAWLEY, 2005 Drosophila: a laboratory handbook. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY.

BAKER, H. G., 1948 Stages in invasion and replacement demonstrated by species of Melandrium. Journal of Ecology **36:** 96-119.

BERGERO, R., A. FORREST, E. KAMAU and D. CHARLESWORTH, 2007 Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: evidence from new sex-linked genes. Genetics **175:** 1945-1954.

BETRAN, E., K. THORNTON and M. LONG, 2002 Retroposed new genes out of the X in Drosophila. Genome Research **12:** 1854-1859.

BROSSEAU, G. E., 1960 Genetic Analysis of the Male Fertility Factors on the Y-Chromosome of *Drosophila melanogaster*. Genetics **45:** 257-274.

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

CHASE, C. D., 2007 Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. Trends in Genetics **23:** 81-90.

CORNELISSEN, T., and P. STILING, 2005 Sex-biased herbivory: a meta-analysis of the effects of gender on plant-herbivore interactions. Oikos **111:** 488-500.

DELPH, L. F., J. L. GEHRING, F. M. FREY, A. M. ARNTZ and M. LEVRI, 2004 Genetic constraints on floral evolution in a sexually dimorphic plant revealed by artificial selection. Evolution **58:** 1936-1946.

DELPH, L. F., F. N. KNAPCZYK and D. R. TAYLOR, 2002 Among-population variation and correlations in sexually dimorphic traits of *Silene latifolia*. Journal of Evolutionary Biology **15:** 1011-1020.

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.

ELLEGREN, H., and J. PARSCH, 2007 The evolution of sex-biased genes and sex-biased gene expression. Nature Reviews Genetics **8:** 689-698.

EMERSON, J. J., H. KAESSMANN, E. BETRAN and M. Y. LONG, 2004 Extensive gene traffic on the mammalian X chromosome. Science **303:** 537-540.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

FILATOV, D. A., 2005 Evolutionary history of *Silene latifola* sex chromosomes revealed by genetic mapping of four genes. Genetics **170:** 975-979.

FILATOV, D. A., 2008 A selective sweep in or near the *Silene latifolia* X-linked gene SlssX. Genetical Research **90:** 85-95.

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.

FU, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915-925.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

GUPTA, V., M. PARISI, D. STURGILL, R. NUTTALL, M. DOCTOLERO *et al.*, 2006 Global analysis of X-chromosome dosage compensation. J Biol **5:** 3.

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.

HOWELL, E. C., S. J. ARMSTRONG and D. A. FILATOV, 2009 Evolution of Neo-Sex Chromosomes in *Silene diclinis*. Genetics **182:** 1109-1115.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159.

ISHIZAKI, K., Y. SHIMIZU-UEDA, S. OKADA, M. YAMAMOTO, M. FUJISAWA *et al.*, 2002 Multicopy genes uniquely amplified in the Y chromosome-specific repeats of the liverwort *Marchantia polymorpha*. Nucleic Acids Research **30:** 4675-4681.

KAISER, V. B., and H. ELLEGREN, 2006 Nonrandom distribution of genes with sex-biased expression in the chicken genome. Evolution **60:** 1945-1951.

KHIL, P. P., N. A. SMIRNOVA, P. J. ROMANIENKO and R. D. CAMERINI-OTERO, 2004 The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. Nature Genetics **36:** 642-646.

KONDO, M., U. HORNUNG, I. NANDA, S. IMAI, T. SASAKI *et al.*, 2006 Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. Genome Research **16:** 815-826.

KRISCHIK, V. A., and R. F. DENNO, 1990 Patterns of growth, reproduction, defense, and herbivory in the dioecious shrub *Baccharis halimifolia* (Compositae). Oecologia **83:** 182-190.

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

LAPORTE, V., D. A. FILATOV, E. KAMAU and D. CHARLESWORTH, 2005 Indirect evidence from DNA sequence diversity for genetic degeneration of the Y-chromosome in dioecious species of the plant Silene: the *SlY4/SlX4* and *DD44-X/DD44-Y* gene pairs. Journal of Evolutionary Biology **18:** 337-347.

LERCHER, M. J., A. O. URRUTIA and L. D. HURST, 2003 Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. Molecular Biology and Evolution **20:** 1113-1116.

LLOYD, D. G., and C. J. WEBB, 1977 Secondary Sex Characters in Plants. Botanical Review **43:** 177-216.

MANK, J. E., and H. ELLEGREN, 2009 Sex-linkage of sexually antagonistic genes is predicted by female, but not male, effects in birds. Evolution **63:** 1464-1472.

MANK, J. E., L. HULTIN-ROSENBERG, M. T. WEBSTER and H. ELLEGREN, 2008 The unique genomic properties of sex-biased genes: Insights from avian microarray data. BMC Genomics **9**.

MATSUNAGA, S., E. ISONO, E. KEJNOVSKY, B. VYSKOT, J. DOLEZEL *et al.*, 2003 Duplicative transfer of a MADS box gene to a plant Y chromosome. Molecular Biology and Evolution **20:** 1062-1069.

MATZ, M., D. SHAGIN, E. BOGDANOVA, O. BRITANOVA, S. LUKYANOV *et al.*, 1999 Amplification of cDNA ends based on template-switching effect and step-out PCR. Nucleic Acids Research **27:** 1558-1560.

MERZOUKI, A., R. T. M'RABET and J. M. MESA, 1996 Sex differentiation of dioecious hemp (*Cannabis sativa* L.) with a precocious morphocinetic character. Archivio Geobotanico **2:** 165-169.

MUELLER, J. L., S. K. MAHADEVAIAH, P. J. PARK, P. E. WARBURTON, D. C. PAGE *et al.*, 2008 The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. Nature Genetics **40:** 794-799.

NAM, K., and H. ELLEGREN, 2008 The Chicken (*Gallus gallus*) Z chromosome contains at least three nonlinear evolutionary strata. Genetics **180:** 1131-1136.

NGUYEN, D. K., and C. M. DISTECHE, 2006 Dosage compensation of the active X chromosome in mammals. Nature Genetics **38:** 47-53.

NICOLAS, M., G. MARAIS, V. HYKELOVA, B. JANOUSEK, V. LAPORTE *et al.*, 2005 A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. Plos Biology **3:** 47-56.

OLIVER, B., and M. PARISI, 2004 Battle of the Xs. Bioessays **26:** 543-548.

PARISI, M., R. NUTTALL, D. NAIMAN, G. BOUFFARD, J. MALLEY *et al.*, 2003 Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science **299:** 697-700.

PRASAD, N. G., and S. BEDHOMME, 2006 Sexual conflict in plants. Journal of Genetics **85:** 161-164.

RICE, W. R., 1984 Sex-chromosomes and the evolution of sexual dimorphism. Evolution **38:** 735-742.

ROSS, M. T., D. V. GRAFHAM, A. J. COFFEY, S. SCHERER, K. MCLAY *et al.*, 2005 The DNA sequence of the human X chromosome. Nature **434:** 325-337.

SATHER, D. N., A. YORK, K. J. POBURSKY and E. M. GOLENBERG, 2005 Sequence evolution and sex-specific expression patterns of the C class floral identity gene, SpAGAMOUS, in dioecious *Spinacia oleracea* L. Planta **222:** 284-292.

STAM, P., 1993 Construction of integrated genetic-linkage maps by means of a new computer package - Joinmap. Plant Journal **3:** 739-744.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

TAJIMA, F., 1990 Relationship between DNA polymorphism and fixation time. Genetics **125:** 447-454.

VEYRUNES, F., P. D. WATERS, P. MIETHKE, W. RENS, D. MCMILLAN *et al.*, 2008 Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Research **18:** 965-973.

VICOSO, B., and B. CHARLESWORTH, 2006 Evolution on the X chromosome: unusual patterns and processes. Nature Reviews Genetics **7:** 645-653.

WALL, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genetical Research **74:** 65-79.

WANG, P. J., J. R. MCCARREY, F. YANG and D. C. PAGE, 2001 An abundance of X-linked genes expressed in spermatogonia. Nature Genetics **27:** 422-426.

WARMKE, H. E., and S. L. J. LEE, 1978 Pollen abortion in T-cytoplasmic male-sterile corn (*Zea mays*) - suggested mechanism. Science **200:** 561-563.

YU, Q., D. STEIGER, E. M. KRAMER, P. H. MOORE and R. MING, 2008 Floral MADS-box genes in trioecious Papaya: characterization of AG and API subfamily genes revealed a

sex-type-specific gene. Tropical Plant Biology **1:** 97-107.

5 High sequence diversity and possible introgression of an X-linked gene on a plant sex chromosome

Atanassov, I., C. Delichere, D. A. Filatov, D. Charlesworth, I. Negrutiu *et al.*, 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., 2008 The temporal dynamics of processes underlying Y chromosome degeneration. Genetics **179:** 1513-1525.

Bachtrog, D., E. Hom, K. M. Wong, X. Maside and P. de Jong, 2008 Genomic degradation of a young Y chromosome in Drosophila miranda. Genome Biology **9:** R30.

Baker, H. G., 1947 Accounts of *Melandrium*, *M. dioicum* and *M. album* for the Biological Flora of the British Isles sponsored by the British Ecological Society. J. Ecol. **35:** 271-292.

Baker, H. G., 1948 Stages in invasion and replacement demonstrated by species of Melandrium. Journal of Ecology **36:** 96-119.

Bergero, R., and D. Charlesworth, 2009 The evolution of restricted recombination in sex chromosomes. Trends in Ecology and Evolution **24:** 94-102.

Bergero, R., D. Charlesworth, D. A. Filatov and R. C. Moore, 2008 Defining regions and rearrangements of the *Silene latifolia* Y chromosome. Genetics **178:** 2045-2053.

Bergero, R., A. Forrest, E. Kamau and D. Charlesworth, 2007 Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: Evidence from new sex-linked genes. Genetics **175:** 1945-1954.

Bopp, S., and G. Gottsberger, 2004 Importance of *Silene latifolia* ssp. *alba* and *S. dioica* (Caryophyllaceae) as host plants of the parasitic pollinator *Hadena bicruris* (Lepidoptera, Noctuidae). Oikos **105:** 221-228.

Cermak, T., Z. Kubat, R. Hobza, A. Koblizkova, A. Widmer *et al.*, 2008 Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes. Chromosome Research **16:** 961-976.

Charlesworth, B., 1978 Model for evolution of Y-chromosomes and dosage compensation. Proceedings of the National Academy of Sciences of the United States of America **75:** 5618-5622.

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

Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver *et al.*, 2007 evolution of genes and genomes on the Drosophila phylogeny. Nature **450:** 203-218.

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.

Filatov, D. A., 2005 Evolutionary history of *Silene latifolia* sex chromosomes revealed by genetic mapping of four genes. Genetics **170:** 975-979.

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.

FU, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915-925.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

GORDO, I., A. NAVARRO and B. CHARLESWORTH, 2002 Muller's ratchet and the pattern of variation at a neutral locus. Genetics **161:** 835-848.

HAAG-LIAUTARD, C., M. DORRIS, X. MASIDE, S. MACASKILL, D. L. HALLIGAN *et al.*, 2007 Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. Nature **445:** 82-85.

HOBZA, R., M. LENGEROVA, J. SVOBODA, H. KUBEKOVA, E. KEJNOVSKY *et al.*, 2006 An accumulation of tandem DNA repeats on the Y chromosome in *Silene latifolia* during early stages of sex chromosome evolution. Chromosoma **115:** 376-382.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159.

IRONSIDE, J. E., and D. A. FILATOV, 2005 Extreme population structure and high interspecific divergence of the Silene Y chromosome. Genetics **171:** 705-713.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The hitchhiking effect revisited. Genetics **123:** 887-899.

KARRENBERG, S., and A. FAVRE, 2008 Genetic and ecological differentiation in the hybridizing campions *Silene dioica* and *S. latifolia*. Evolution **62:** 763-773.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings in Bioinformatics **5:** 150-163.

LAPORTE, V., and B. CHARLESWORTH, 2002 Effective population size and population subdivision in demographically structured populations. Genetics **162:** 501-519.

LAPORTE, V., D. A. FILATOV, E. KAMAU and D. CHARLESWORTH, 2005 Indirect evidence from DNA sequence diversity for genetic degeneration of the Y-chromosome in dioecious species of the plant Silene: the *SlY4/SlX4* and *DD44-X/DD44-Y* gene pairs. Journal of Evolutionary Biology **18:** 337-347.

MARAIS, G. A. B., M. NICOLAS, R. BERGERO, P. CHAMBRIER, E. KEJNOVSKY *et al.*, 2008 Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. Current Biology **18:** 545-549.

MATSUNAGA, S., E. ISONO, E. KEJNOVSKY, B. VYSKOT, J. DOLEZEL *et al.*, 2003 Duplicative transfer of a MADS box gene to a plant Y chromosome. Molecular Biology and Evolution **20:** 1062-1069.

MAYNARD SMITH, J., and J. HAIGH, 1974 Hitch-hiking effect of a favorable gene. Genetical Research **23:** 23-35.

MINDER, A. M., C. ROTHENBUEHLER and A. WIDMER, 2007 Genetic structure of hybrid zones between *Silene latifolia* and *Silene dioica* (Caryophyllaceae): evidence for introgressive hybridization. Molecular Ecology **16:** 2504-2516.

MINDER, A. M., and A. WIDMER, 2008 A population genomic analysis of species boundaries: neutral processes, adaptive divergence and introgression between two hybridizing plant species. Molecular Ecology **17:** 1552-1563.

MULLER, H. J., 1964 The relation of recombination to mutational advance. Mutation Research **1:** 2-9.

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.

NICOLAS, M., G. MARAIS, V. HYKELOVA, B. JANOUSEK, V. LAPORTE *et al.*, 2005 A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. Plos Biology **3:** 47-56.

PRENTICE, H. C., J. U. MALM and L. HATHAWAY, 2008 Chloroplast DNA variation in the European herb *Silene dioica* (red campion): postglacial migration and interspecific introgression. Plant Systematics and Evolution **272:** 23-37.

RAHME, J., A. WIDMER and S. KARRENBERG, 2009 Pollen competition as an asymmetric reproductive barrier between two closely related Silene species. Journal of Evolutionary Biology **22:** 1937-1943.

REICH, D. E., S. F. SCHAFFNER, M. J. DALY, G. MCVEAN, J. C. MULLIKIN *et al.*, 2002 Human genome sequence variation and the influence of gene history, mutation and recombination. Nature Genetics **32:** 135-142.

SACHIDANANDAM, R., D. WEISSMAN, S. C. SCHMIDT, J. M. KAKOL, L. D. STEIN *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928-933.

STROBECK, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117:** 149-153.

SWEIGART, A. L., and J. H. WILLIS, 2003 Patterns of nucleotide diversity in two species of Mimulus are affected by mating system and asymmetric introgression. Evolution **57:** 2490-2506.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

WOLFE, K. H., P. M. SHARP and W. H. LI, 1989 Mutation-rates differ among regions of the mammalian genome. Nature **337:** 283-285.

ZLUVOVA J, GEORGIEV S, JANOUSEK B, CHARLESWORTH D, VYSKOT B *et al.*, 2007 Early events in the evolution of the *Silene latifolia* Y chromosome: male specialization and recombination arrest. Genetics **177:** 375-386.

ZLUVOVA, J., B. JANOUSEK, I. NEGRUTIU and B. VYSKOT, 2005 Comparison of the X and Y Chromosome Organization in *Silene latifolia*
Genetics **170:** 1431 - 1434.

ATANASSOV, I., C. DELICHERE, D. A. FILATOV, D. CHARLESWORTH, I. NEGRUTIU *et al.*, 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., 2004 Evidence that positive selection drives Y-chromosome degeneration in *Drosophila miranda*. Nature Genetics **36:** 518-522.

BETANCOURT, A. J., J. J. WELCH and B. CHARLESWORTH, 2009 Reduced Effectiveness of Selection Caused by a Lack of Recombination. Current Biology **19:** 655-660.

CARVALHO, A. B., L. B. KOERICH and A. G. CLARK, 2009 Origin and evolution of Y chromosomes: Drosophila tales. Trends in Genetics **25:** 270-277.

CROW, J. F., and M. J. SIMMONS, 1983 The mutation load in Drosophila. In: The genetics and biology of Drosophila. Academic Press, London.

DELICHERE, C., J. VEUSKENS, M. HERNOULD, N. BARBACAR, A. MOURAS *et al.*, 1999 *SlY1*, the first active gene cloned from a plant Y chromosome, encodes a WD-repeat protein. Embo Journal **18:** 4169-4179.

FILATOV, D. A., 2008 A selective sweep in or near the *Silene latifolia* X-linked gene *SlssX*. Genetical Research **90:** 85-95.

HONYS, D., and D. TWELL, 2003 Comparative analysis of the Arabidopsis pollen transcriptome. Plant Physiology **132:** 640-652.

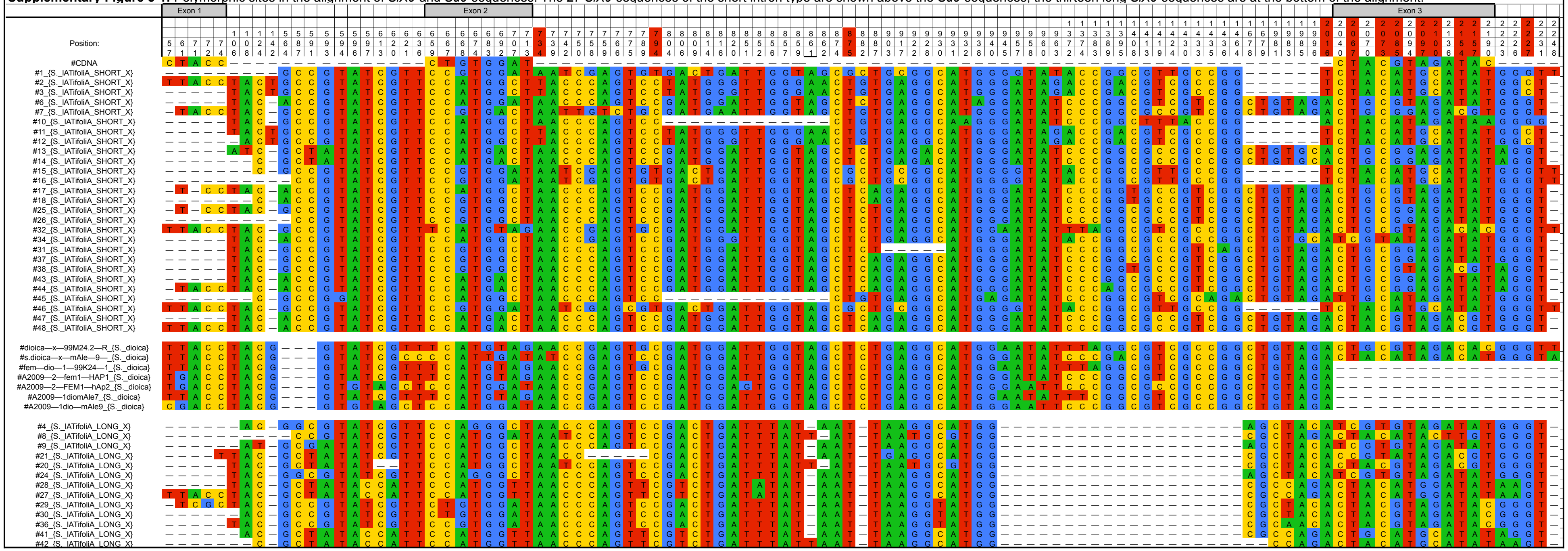
MOORE, R. C., O. KOZYREVA, S. LEBEL-HARDENACK, J. SIROKY, R. HOBZA *et al.*, 2003 Genetic and functional analysis of *DD44*, a sex-linked gene from the dioecious plant *Silene latifolia*, provides clues to early events in sex chromosome evolution. Genetics **163:** 321-334.

NORDBORG, M., B. CHARLESWORTH and D. CHARLESWORTH, 1996 The effect of recombination on background selection. Genetical Research **67:** 159-174.

SKALETSKY, H., T. KURODA-KAWAGUCHI, P. J. MINX, H. S. CORDUM, L. HILLIER *et al.*, 2003 The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423:** 825-U822.

8 Supplementary Material

Supplementary Figure 5-1: Polymorphic sites in the alignment of *SlX9* and *Sd9* sequences. The 27 *SlX9* sequences of the short intron type are shown above the *Sd9* sequences, the thirteen long *SlX9* sequences are at the bottom of the alignment.


Exon 1
Exon 2
Exon 3
Position:
#CDNA
#1_{S._lATifoliA_SHORT_X}
#2_{S._lATifoliA_SHORT_X}
#3_{S._lATifoliA_SHORT_X}
#6_{S._lATifoliA_SHORT_X}
#7_{S._lATifoliA_SHORT_X}
#10_{S._lATifoliA_SHORT_X}
#11_{S._lATifoliA_SHORT_X}
#12_{S._lATifoliA_SHORT_X}
#13_{S._lATifoliA_SHORT_X}
#14_{S._lATifoliA_SHORT_X}
#15_{S._lATifoliA_SHORT_X}
#16_{S._lATifoliA_SHORT_X}
#17_{S._lATifoliA_SHORT_X}
#18_{S._lATifoliA_SHORT_X}
#25_{S._lATifoliA_SHORT_X}
#26_{S._lATifoliA_SHORT_X}
#32_{S._lATifoliA_SHORT_X}
#34_{S._lATifoliA_SHORT_X}
#31_{S._lATifoliA_SHORT_X}
#37_{S._lATifoliA_SHORT_X}
#38_{S._lATifoliA_SHORT_X}
#43_{S._lATifoliA_SHORT_X}
#44_{S._lATifoliA_SHORT_X}
#45_{S._lATifoliA_SHORT_X}
#46_{S._lATifoliA_SHORT_X}
#47_{S._lATifoliA_SHORT_X}
#48_{S._lATifoliA_SHORT_X}
#dioica—x—99M24.2—R_{S._dioica}
#s.dioica—x—mAle—9—_{S._dioica}
#fem—dio—1—99K24—1_{S._dioica}
#A2009—2—fem1—HAP1_{S._dioica}
#A2009—2—FEM1—hAp2_{S._dioica}
#A2009—1diomAle7_{S._dioica}
#A2009—1dio—mAle9_{S._dioica}
#4_{S._lATifoliA_LONG_X}
#8_{S._lATifoliA_LONG_X}
#9_{S._lATifoliA_LONG_X}
#21_{S._lATifoliA_LONG_X}
#20_{S._lATifoliA_LONG_X}
#24_{S._lATifoliA_LONG_X}
#28_{S._lATifoliA_LONG_X}
#27_{S._lATifoliA_LONG_X}
#29_{S._lATifoliA_LONG_X}
#30_{S._lATifoliA_LONG_X}
#36_{S._lATifoliA_LONG_X}
#41_{S._lATifoliA_LONG_X}
#42_{S._lATifoliA_LONG_X}

```
           PROGRAM projectrec2
         IMPLICIT NONE

         INTEGER, PARAMETER :: nZyg =1000, nGam = 1, nChr=100

         INTEGER, DIMENSION (nZyg, nGam, nChr):: n1, n0 , n2,n4,  m1, m2, g
         INTEGER, DIMENSION (2, 1, nChr)::embryo, tmp
         INTEGER:: donor, aceptor
         INTEGER, DIMENSION(1:2):: par
         INTEGER:: iZyg, iGam, iChr, boo1, iLoc, maske, ix, i, gen, iff, j,

         INTEGER:: NoMut, n, newZyg
         REAL:: v,a,b,c
         REAL::nome, pRec, lambda, critfit
         INTEGER:: Ce
         INTEGER:: aRec, test, st, chrom, bits
         INTEGER:: ia,ja, Ne, k, homoMask, heteroMask, homoTest, heterotest
         REAL:: w,x, y,  homofitness, heterofitness, homofitness_so_far, he
         REAL, Dimension(1:nZyg)::fitness,relative_fitness, ln_fitness, rel
        REAL, DIMENSION(0:31, nChr)::s
        INTEGER, DIMENSION(0:31)::pos
         INTEGER:: numg, nPRINT, try, words, doloops, total, itotal, conv
         PARAMETER (numg=32)
        CHARACTER:: out_file
       REAL:: harmonic_s,  harmonic_summe, harmonic_mean, harmonic_mean_Nes
           REAL:: summe, mean, meanNes


         REAL:: lambdarealnochr,rhalfl, realnochr
         INTEGER::  NoConv, noChrright, noChrleft, reset , halfl, rest, unt
         INTEGER::  iGamotfher , Chr, cChr, locus, noRec, norecs, il, ipos,
        REAL:: average_fitness,  lambdapRec, noms, vei, fitness_term, T2, T
         REAL:: freq_average_1, frequ1, frequ2, frequ3, freq_average_0, no_
          REAL:: AVERAGE_FITNESSRELATIVE, var, vari, fitness_var, coef_of_v
                 REAL:: sum_coeff, count_sum, average_sel_coeff



     n2=0

 doloops =4444
       DO i = 1,doloops
       Call random_number(v)
       END DO


        DO iCHr = 1,nChr
```

```
         !determine selection coefficients for positions 0 -31
         Ne = 500 !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!needs to
         fitness = 1.0
       DO ja = 0,31
         w = 0
                 DO ia = 1, 12
                 CALL RANDOM_Number(v)
                 w =w+v  ! w = Sum of 12 random nos., N(0,1) distributed
                 END DO

         w = w-6.0
         x = -3.3+(w*1.106)      ! 0.00142 rate of lethal mut.
         y = exp(x)              ! lognormal distribtuon of y

         IF (y .le. 1.0) THEN
         s(ja, iChr)  = y
         ELSE
         s(ja, iChr) = 1.0

         END IF

       END DO

         !positions on the chromosome:
         DO ja = 0,31
       k = 2**ja
       pos(ja) = k
      ! PRINT*, pos(ja)
       END DO

  END DO


  st = 3
  Do chrom = 1,3,2
  st = st-1
  DO  iChr = chrom,nChr,3        !make selection coeff. at every third pos.

       DO ja = st,31,3
       s(ja, iChr) = 0.0
       END DO
 END DO
 END DO

 DO iChr = 2,nChr,3
 DO ja = 0 ,31,3
       s(ja, iChr) = 0.0
       END DO
       END DO
```

```fortran
 PRINT*, "check"
!Calculate arithmetic mean selection coeff
summe = 0.0
harmonic_summe = 0.0
counts = 0.0

Do iChr = 1,nChr
Do iPos = 0,31
!PRINT*, " "
!PRINT*, "s", s(iPos, iChr)

If( ( s(iPos, iChr).ne. 1.0) .and. ( s(iPos, iChr).ne. 0.0) ) THEN
counts = counts + 1.0
harmonic_s = 1.0/(s(iPos, iChr))                !=1/s
!PRINT*, "harmonic_s", harmonic_s
summe = summe + s(iPos, iChr)
harmonic_summe = harmonic_summe + harmonic_s    ! = sum(1/s)
!PRINT*, "harmonic_summe", harmonic_summe
ENDIF

ENDDO
ENDDO
!PRINT*, counts, "counts"
mean = summe/counts
PRINT*, " arithmetic mean selection coeff.:", mean
meanNes = mean*nZyg*0.5
PRINT*, " arithmetic mean Nes", meanNes

harmonic_mean = counts/harmonic_summe
PRINT*, harmonic_mean, "harmonic_mean"
harmonic_mean_Nes = harmonic_mean*nZyg*0.5


PRINT*, "harmonic_mean_Nes scaled down to haploids", harmonic_mean_Nes


n2 = 0


 CALL set_selected(n2,s,nZyg,nGam,nChr)
!PRINT*, n2
 CALL set_neutral(n2,nZyg,nGam,nChr)
 PRINT*, maxval (n2)
 PRINT*, minval (n2)

!WRITE (UNit= 1, fmt = "(I20)") n2
!WRITE (unit = 2, fmt = "(F10.8)") s
 PRINT*, "selection and n2 set"
```

```
lethal = 0
Do iChr = 1, nChr
Do iPos = 0,31
IF (s(iPos, iChr)== 1.0) THEN
lethal = lethal +1
ENDIF
ENDDO
ENDDO
PRINT*, "lethals" , lethal

conv = 0
 doloops =4444
      DO i = 1,doloops
      Call random_number(v)
      END DO
 n4 = 0
T2=0
total = 0
itotal=0
fitness = 1.0
 relative_fitness = 1.0

equil_fitness=exp(-32*nChr*0.0000104*0.6666666) !w = exp(U) !changed from

lambdapRec = 0.000026*nChr*32!!keeps Ner = 10-8 * 1.3 * 10^6constant!!!adj
 recfrequ= 0.000026     !!!!!!!!!!!!!!!!!!!
        GCfrequ = 0.000018466   !!!!!!!!!!!!!!!!!!!!!!!!!
        lambdaGC =0.000018466*nChr*32   !!!!!!!!!!!!!!!0.25*10-5/352*e con
        !lambdapRec=0.0
        !recfrequ=0.0
        !GCfrequ=0.0
        !lambdaGC = 0.0




nPrint = 1000
!PRINT*, "nPrint", nPrint
average_fitness = 1.0
          n0 = 0
nomutations = 0


PRINT*, "nChr", nChr
PRINT*, "GCfrequ", GCfrequ
PRINT*, " recfrequ",  recfrequ


!open( unit = 3, File = "n2-after_sel-haploids-20000-4th-NORECGC-newfitnes
!OPEN(unit = 5, file = "ntdiv-selection10,000-2nd-RECCGC", status = "new")
!OPEN(unit = 7, file = "fitness-selection-haploids-10000-noRECGC", status =
!PRINT*, n2
!PRINT*, s

 !positions on the chromosome:
        DO ja = 0,31
      k = 2**ja
      pos(ja) = k
      END DO



!CALL fixedsyn(n2, nZyg, nGam, nChr)
!call ntdiv_neutral(n2, nZyg, nGam, nChr)

!!!!!!!!!DETERMINE FITNESS


Do iZyg = 1,nZyg

homofitness_so_far = 0.0

chromosomess:   DO iChr = 1,nChr

        homofitness = 0.0



  cx1:      Do ja = 0,31
  fitness_term = 0.0
        homoTest = iand(n2(iZyg, 1,iChr), pos(ja))
                IF (homotest.NE.0) THEN

                If ( s(ja,iChr) == 0.0 ) THEN
                fitness_term = 0.0
                ELSEIF ( s(ja, iChr ) .ge. 1.0)  THEN           !lethal
                !PRINT*, "homolethal", iZyg
                homofitness_so_far=-10000.0

                EXIT Chromosomess
                ELSE
                fitness_term =log(1.0-(s(ja, iChr)))
                !PRINT*, izyg, fitness_term
                ENDIF
        ENDIF
        homofitness=homofitness+fitness_term    !Multiplicative fitness ef

        END DO cx1

        homofitness_so_far = homofitness_so_far+homofitness


         END DO Chromosomess


IF ( homofitness_so_far == -10000.0 ) THEN
relative_fitness_compare(iZyg)  = -10000000.0


ELSE

ln_fitness(iZyg) = homofitness_so_far
relative_fitness_compare(iZyg) =ln_fitness(iZyg)-log(equil_fitness)     !th
ENDIF


ENDDO



Do iZyg = 1, nZyg
If (relative_fitness_compare(iZyg)  == -10000000.0) THEN
 relative_fitness(iZyg) = 0.0
 ELSE
   relative_fitness(iZyg) = exp(relative_fitness_compare(iZyg)-maxval(rela
   ENDIF
   ENDDO





CALL ntdiv_neutral(n2,nZyg, nGam, nChr)
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

PRINT*, maxval(s), "MAXVAL S"
        !Number of mutations from Poisson distibution and Locus of mutatio
generation:     DO gen = 1,10000
     !  PRINT*, "GENERATION", gen

 Zyg:           DO iZyg = 1, nZyg

                DO iGam = 1, nGam
        Chro:           DO iChr = 1, nChr
                        NoMut = 0
                         CALL RANDOM_NUMBER(v)
                                nome = v
                                lambda = 0.0003328!0.0003328!corresp. to m
                                CALL poissontry(nome, Ce,lambda)        !

        mutations:              IF (Ce == 0) THEN  !no mutations

                        n2(iZyg, iGam, iChr) = n2(iZyg, iGam, iChr)

                        n0(iZyg, iGam, iChr) = n2(iZyg, iGam, iChr)


                        ELSeIF  (Ce>0) THEN mutations

                                DO  WHILE (NoMut .lt.  Ce)
                                nomutations = nomutations+1
                                 CALL RANDOM_NUMBER(v)    ! random positio
                                n=v*32
                                bits = btest1(n2(iZyg, iGam, iChr), n)

                                IF (bits == 1) THEN
                                n2(iZyg, iGam, iChr) = ibclr(n2(iZyg, iGam
                                ELSEIF (bits == 0) THEN

                                n1(iZyg, iGam, iChr) = 2**n
                                n2(iZyg, iGam, iChr) = ior(n2(iZyg, iGam,
                                n0(iZyg, iGam, iChr) = n2(iZyg, iGam, iChr
                                END IF

                                NoMut = NoMut + 1

                                END Do

                                ENDIF mutations

                        END DO Chro

                END DO

END DO Zyg

 !!!!!!!!!!DETERMINE FITNESS


Do iZyg = 1,nZyg
```

```
homofitness_so_far = 0.0

chromosomess11: DO iChr = 1,nChr

        homofitness = 0.0



  cx11:      Do ja = 0,31
  fitness_term = 0.0
        homoTest = iand(n2(iZyg, 1,iChr), pos(ja))
                IF (homotest.NE.0) THEN

                If ( s(ja,iChr) == 0.0 ) THEN
                fitness_term = 0.0
                ELSEIF ( s(ja, iChr ) .ge. 1.0)  THEN           !lethal
                !PRINT*, "homolethal", iZyg
                homofitness_so_far=-10000.0

                EXIT Chromosomess11
                ELSE
                fitness_term =log(1.0-(s(ja, iChr)))
                !PRINT*, izyg, fitness_term
                ENDIF

        ENDIF
        homofitness=homofitness+fitness_term    !Multiplicative fitness ef

        END DO cx11

        homofitness_so_far = homofitness_so_far+homofitness


         END DO Chromosomess11


!PRINT*, equil_fitness, "eqil_fitness"

IF ( homofitness_so_far == -10000.0 ) THEN
relative_fitness_compare(iZyg)  = -10000000.0


ELSE

ln_fitness(iZyg) = homofitness_so_far
relative_fitness_compare(iZyg) =ln_fitness(iZyg)-log(equil_fitness)     !t
ENDIF
```

```
ENDDO



Do iZyg = 1, nZyg
If (relative_fitness_compare(iZyg)  == -10000000.0) THEN
 relative_fitness(iZyg) = 0.0
 ELSE
   relative_fitness(iZyg) = exp(relative_fitness_compare(iZyg)-maxval(rela
   ENDIF

   ENDDO



        !CHOOSE PARENTS

      newZyg = 1

counting:        Do  WHILE (newZyg .le. nZyg) !ends when as many viable zy


sampling1:              DO

        CALL RANDOM_NUMBER(v)
        par(1) = v*nZyg+1

        CALL RANDOM_number(v)
        IF (v .lt. relative_fitness(par(1)) ) Exit sampling1
        ENDDO sampling1

  sampling2:      DO

        CALL RANDOM_NUMBER(v)
        par(2) = v*nZyg+1
        CALL RANDOM_number(v)
        IF (v .lt. relative_fitness(par(2)) ) Exit sampling2

        ENDDO sampling2



Do iChr = 1, nChr

embryo(1,1,iChr) = n2(par(1), 1, iChr)

embryo(2,1,iChr) = n2(par(2), 1, iChr)

ENDDO
```

```
!!!!!!!!!!!!!!!!!!
!gene conversion
!!!!!!!!!!!!!!!!
 IF (lambdaGC .ne. 0) THEN

         CALL RANDOM_NUMBER(v)
         nome = v

        CALL poissontry(nome, Ce,lambdaGC)      ! number of gene conversio
        NoConv = 0

        IF (Ce ==0) THEN
        !PRINT*, "no gene conversion"
        ELSE

        DO WHILE (NoConv .lt. Ce)
conv = conv+1
        Call random_number(v)
        cChr = v*nChr +1        !chromosome for gene conversion

        CALL random_number(v)   !Locus from even distribution
        locus=v*32

   iGam = 1


 !  decide which gamete donor:
     Call Random_number(v)
        IF (v .lt. 0.5) THEN
       donor = 1
      aceptor =2
        ELSE
       donor = 2
      aceptor =1
        END IF




        Call Random_number(v)
               halfl =  (tractlength(v))/2      !integer half tract length
        rhalfl=real(halfl)
        realnoChr= rhalfl/32.0
         noChrright = realnoChr !this is the number of chromosomes that ar
       noChrleft = realnoChr

       rest = (realnoChr-real(noChrright))*32
        !for right hand side of conversion locus:
```

```
        untilright = locus + rest
        if (untilright .gt. 31 )THEN
        untilright = untilright -32
        ENDIF

        IF ((noChrright ==0) .and. (untilright.ge.locus)) then
        noChrright = -1
        ELSEIF ((noChrright == 0) .and. (untilright .lt. locus)) THEN
        noChrright = 0
        ELSEIF ((nochrright .ge. 1) .and. (untilright .ge. locus)) ThEN
        noChrright = noCHRRIGHT -1
        ELSEIF ((NOCHRRIGHT .ge. 1) .AND. (UNTILRIGHT .lt. LOCUS)) THEN
        noChrright = noChrright
        ENDIF
           !for left hand side of conversion locus:

         untilleft = locus - rest

         IF( untilleft .lt. 0 )THEN
         untilleft = 32-abs(untilleft)
         ENDIF

         IF ((noChrleft ==0) .and.  (untilleft .le. locus)) THEN
         noCHRleft = -1
         ELSEIF ((noChrleft==0) .and. (untilleft .gt. locus) )THEN
         noChrleft = 0
         ELSEIF ((noChrleft .ge. 1) .and. (untilleft .gt. locus )) THEN
         noChrleft = noChrleft
         ELSEIF ((noChrleft .ge.1) .and. (untilleft .le. locus )) THEN
         noChrleft = noChrleft -1
         ENDIF

         IF (noChrleft == -1) THEN
         call mvbits(embryo(donor, 1, cChr), untilleft, locus+1-untilleft,
         ELSE
         call mvbits(embryo(donor, 1, cChr), 0, locus+1, embryo(aceptor, 1,
         ENDIF

         IF ((noChrleft .ge.0) .and. ( cChr - noChrleft .ge.2))THEN  !!!!!!
         call mvbits(embryo(donor, 1, cChr-noChrleft-1), untilleft, 32-unti
         ENDIF


          !chromosomes fully converted:

          IF ((noChrleft .ge.1) .and. (noChrright.ge.1) .and.((cChr-noChrle
         DO Chr = cChr-noChrleft, cChr+noChrright        !all to be convert
         call mvbits(embryo(donor, 1, Chr), 0, 32, embryo(aceptor, 1, Chr),
         END DO


        ELSEIF  ((noChrleft .ge.1) .and. (noChrright.ge.1) .and.((cChr-noChr
        DO Chr = 1,nChr  !all converted
        call mvbits(embryo(donor, 1,  Chr), 0, 32, embryo(aceptor, 1, Chr),
         END DO

        ELSEIF ((noChrleft .ge.1) .and. (noChrright.ge.1) .and.((cChr-noChr
         DO Chr = 1,nChr
        call mvbits(embryo(donor, 1, Chr), 0, 32, embryo(aceptor, 1, Chr), (
         END DO

         ELSEIF ((noChrleft .ge.1) .and. (noChrright.ge.1) .and.((cChr-noChr
         DO Chr = 1,nChr
        call mvbits(embryo(donor, 1,  Chr), 0, 32,embryo(aceptor, 1, Chr), (
         END DO

         ELSEIF ((noChrright == 1) .and. (noChrleft ==0) .and. ((cChr+noChr
         call mvbits(embryo(donor, 1,  cChr+1), 0, 32, embryo(aceptor, 1, c(


         ELSEIF ((noChrleft == 1) .and. (noChrright == 0) .and. ((cChr-noChr
         call mvbits(embryo(donor, 1,  cChr-1), 0, 32, embryo(aceptor, 1, c(
         ENDIF

      ! positions to the right converted

         IF (noChrright == -1) THEN
         call mvbits(embryo(donor, 1,  cChr), locus+1, untilright-locus, emb

         ELSE
         call mvbits(embryo(donor, 1,  cChr), locus+1, 31-locus, embryo(acep
         END IF

         IF ((noChrright .ge. 0) .and. (cChr+ noChrright .le. nChr-1)) THEN
         call mvbits(embryo(donor, 1,  (cChr+noChrright+1)), 0, untilright+
         ENDIF


     NoConv = NoConv +1

        END DO
         ENDIF




 ENDIF
```

```
  !!!!!!!!!!!!!!!!!!
!  END GENE CONVERSION
  !!!!!!!!!!!!!!!!!!!


!!!!!!!!!!!!!!!!!
!RECOMBINATION
!!!!!!!!!!!!!!!!!

      boo1 = 4294967295


!!!!!FIRST RECOMBINATION WITHOUT FORMING GAMETES:

    !chosen parents make recombination & gametes
        itotal=itotal+1

 CALL RANDOM_NUMBER(v)
         nome = v
        CALL poissontry(nome, Ce,lambdapRec)
        !!!!!!!!!!!!!!!!NEW PART:
RECO : If (Ce == 0 .or. lambdapRec == 0.0) THEN
!NO RECOMBINATION
Call random_number(v)

IF(v.ge.0.5) THEN

DO iff = 1,nChr
n4(newZyg, 1,iff) = embryo(1,1,iff)
ENDDO

Else
Do iff = 1,nChr
n4(newZyg, 1,iff) = embryo(2,1,iff)
ENDDO
ENDIF

ELSE RECO
!RECOMBINATION

recDo:          Do il = 1, Ce
 total = total+1

                                Call Random_Number(v)
                                iChr = v*nChr+1


                                 CALL RANDOM_NUMBER(v)
                                iLoc=v*32


recomb_within:          IF (iLoc .ne. 0)        then

        tmp = embryo

      call mvbits (embryo(1,1,iChr), iLoc, 32-iLoc, tmp(2,1,iChr), iLoc)

      call mvbits (embryo(2,1,iChr), iLoc, 32-iLoc, tmp(1,1,iChr), iLoc)


   IF (iChr .ne. nChr) THEN
        DO iff= iChr+1, nChr
        tmp(1,1,iff) = embryo(2,1,iff)
        tmp(2,1,iff) = embryo(1,1,iff)
        END DO
        ENDIF


ELSE recomb_within  !recombination between words
        DO iff= iChr, nChr
        tmp(1,1,iff) = embryo(2,1,iff)
        tmp(2,1,iff) = embryo(1,1,iff)
        END DO
END IF recomb_within


 ENDDO recDo


       CALL RANDOM_NUMBER(v)
     IF (v .lt. 0.5) THEN

        Do i = 1,nChr
        n4(newZyg, 1, i) = tmp(1,1,i)
        END DO

        ELSE
        Do i = 1,nChr
        n4(newZyg, 1, i) = tmp(2,1,i)
        END DO

        ENDIF

        ENDIF RECO
        !PRINT*, "total recs after", newZyg, total
        newZyg = newZyg +1

        END DO COUNTING


        Do iZyg = 1, nZyg
        Do iGam = 1, nGam
        Do iChr = 1, nChr
        n2(iZyg, iGam, iChr) = n4(iZyg, iGam, iChr)
        END DO
        END DO
        END DO

        if (mod(gen,nPrint).eq.0) then

PRINT*, gen


Do iZyg = 1, nZyg
!  PRINT*," fitness(iZyg)", fitness(iZyg)

   ENDDO

!CALL fixedsyn(n2, nZyg, nGam, nChr)  ! PRINTS number of sites fixed for 0
!WRITE(unit = 7, fmt = "(E14.3)") average_fitness
CALL ntdiv_neutral(n2,nZyg, nGam, nChr)

!PRINT*, average_fitness
  Call CPU_time(T3)
      T_int = T3-T2
PRINT*, "CPU time", T_int

      T2=T3

ENDIF


!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
END DO generation
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

PRINT*, "no gene conversions", conv



PRINT*, "nChr", nChr
PRINT*, "GCfrequ", GCfrequ
PRINT*, " recfrequ",  recfrequ
```

```
average_rel_fitness= sum(relative_fitness)/nZyg

var = 0.0

Do iZyg = 1, nZyg

VARI = (relative_fitness(iZyg)-average_rel_fitness)**2
Var = var +vari
ENDDO

fitness_var = Var /(nZyg-1)

coef_of_var = (SQRT(fitness_var))/average_rel_fitness

PRINT*, "coefficient of var. in fitness (relative) with Method (0,1)"
PRINT*,  coef_of_var


CALL Bsel(n2,nZyg, nGam, nChr, s, recfrequ, GCfrequ)

CALL  fixednonsyn(n2, nZyg, nGam, nChr, s) ! PRINTS number of sites fixed
PRINT*, " "

CALL fixedsyn(n2, nZyg, nGam, nChr)  ! PRINTS number of sites fixed for 0

CALL LD_selected(n2,nZyg,nGam,nChr)

CALL LD_neutral(n2,nZyg,nGam,nChr)

PRINT*, "NEW METHOd all sites"


no_test=0.0
no_test_1 = 0.0

!test prop. of nonsyn sites that carry 1

DO iChr = 1,10,3

DO iPos = 0,30,3
IF (s(ipos, iChr)== 0) THEN
PRINT*, "s alarm!", iChr, ipos
ENDIF

Do iZyg = 1,20
Do iGam = 1,nGam
```

```
test= btest1(n2(iZyg, iGam,iChr),iPos)



IF (test== 1) THEn
no_test_1 = no_test_1+1.0
ENDIF
EnD DO
END DO
END DO

DO iPos = 1,31,3
IF (s(ipos, iChr)== 0) THEN
PRINT*, "s alarm!", iChr, ipos
ENDIF

Do iZyg = 1,20
Do iGam = 1,nGam


test= btest1(n2(iZyg, iGam,iChr),iPos)

IF (test== 1) THEn
no_test_1 = no_test_1+1.0
ENDIF
EnD DO
END DO
END DO

END DO


!!!!!!!!CHROMOSOME 2 etc.


DO iChr = 2,10,3

DO iPos = 1,31,3
IF (s(ipos, iChr)== 0) THEN
PRINT*, "s alarm!", iChr, ipos
ENDIF

Do iZyg = 1,20
Do iGam = 1,nGam


test= btest1(n2(iZyg, iGam,iChr),iPos)
```

```fortran
IF (test== 1) THEn
no_test_1 = no_test_1+1.0
ENDIF


EnD DO
END DO
END DO

DO iPos = 2,29,3
Do iZyg = 1,20
Do iGam = 1,nGam


test= btest1(n2(iZyg, iGam,iChr),iPos)
IF (s(ipos, iChr)== 0) THEN
PRINT*, "s alarm!"
ENDIF

IF (test== 1) THEn
no_test_1 = no_test_1+1.0
ENDIF
EnD DO
END DO
END DO

END DO

!FCHROMOSOM 3 etc.


third: DO iChr = 3,10,3

thirdpos: DO iPos = 0,30,3
IF (s(ipos, iChr)== 0) THEN
PRINT*, "s alarm!", iChr, ipos
ENDIF

Do iZyg = 1,20
Do iGam = 1,nGam

test= btest1(n2(iZyg, iGam,iChr),iPos)

IF (test== 1) THEn
no_test_1 = no_test_1+1.0
ENDIF
EnD DO
END DO
END DO thirdpos
```

```
DO iPos = 2,29,3
IF (s(ipos, iChr)== 0) THEN
PRINT*, "s alarm!", iChr, ipos
ENDIF

Do iZyg = 1,20
Do iGam = 1,nGam


test= btest1(n2(iZyg, iGam,iChr),iPos)

IF (test== 1) THEn
no_test_1 = no_test_1+1.0
ENDIF
EnD DO
END DO
END DO

END DO Third


divider = 32.0*10.0*20.0*0.666666
PRINT*, "divider", divider
freq_average_1= no_test_1/divider
PRINT*, "no_test_1"
PRINT*,  no_test_1
PRINT*, "freq_average_1 at nonsyn. sites with new method 10chr, 20zyg site
PRINT*, freq_average_1
freq_average_0 = 1-freq_average_1
PRINT*, " "
PRINT*, "freq_average_0 at nonsyn. sites"
PRINT*,  freq_average_0

PRINT*, "pRec = ", pRec



 !PRINT*, "JUST 2 ZYGOTES"
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!Subroutine
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
    CONTAINS


        SUBROUTINE ntdiv_neutral(n, mZyg, mGam, mChr)
     INTEGER, INTENT(IN):: mZyg, mGam, mChr
     INTEGER:: jChr,  jZyg, jGam,  i, first_test, start, chromosome
     INTEGER, DIMENSION(mZyg, mGam, mChr),  INTENT(IN) ::n
     REAL:: prop1, prop0,propsegr_syn, nositessyn1, nositessyn0,no_fixedsy
     INTEGeR::  BB,AA,  bits1, bits2, ipos,  mult
      REAL::CC,F

     REAL:: nuc
    INTEGER::      test, no_segregating, ende
     REAL :: theta
     REAL:: a,b,nucleotidediv
     REAL:: a1,a2,b1,b2,c1,c2,e1,e2, S, D, a11, btes, n_2, t, t1, t2, var,





     no_fixedsyn0 = 0.0
        no_fixedsyn1 = 0.0
          CC = 0
        start = 4
     DO chromosome = 1,2
     start = start - 2
     DO jChr = chromosome,mChr,3
     Do i = start,31,3

          !!!DIVERsiTY!!!!!!!!!!!!!!!!!!!
       BB = 0

Zygote:        DO jZyg = 1, mZyg
!PRINT*, "zyg", jZyg
        AA = 0

        bits1 = btest1(n(jZyg,1,jChr), i)

        AA = bits1
       ! PRINT*, "A", A


        BB = BB+AA              !sum of all "1" at that position

        mult = (mZyg-BB)*BB  !no of pairwise diff. at that pos.

        END DO Zygote


        CC = CC + mult
!!!END DIVERITY!!!!!!!!!!!!!!!!!!


             END DO
        END DO
        END Do

!third chromsome


     DO jChr = 3,mChr,3
     Do i = 1,31,3
         !!!DIVERsiTY!!!!!!!!!!!!!!!!!!!!
       BB = 0

Zygoten:        DO jZyg = 1, mZyg

        AA = 0

        bits1 = btest1(n(jZyg,1,jChr), i)

        AA = bits1



        BB = BB+AA              !sum of all "1" at that position

        mult = (mZyg-BB)*BB  !no of pairwise diff. at that pos.

        END DO Zygoten

        CC = CC + mult
!!!END DIVERITY!!!!!!!!!!!!!!!!!!!!



     END DO
        END DO

 F = mZyg*(mZyg - 1.0)*0.5    !!!!!!! = n(n-1)/2

        nuc = CC/F

nucleotidediv=nuc/(mChr*32*0.333333333)

! WRITE(unit = 5, fmt ="(F10.8)")  nucleotidediv
 PRINT*, nucleotidediv

  END SUBROUTINE ntdiv_neutral
```

```
          SUBROUTINE poissontry(nom, C,lam)

         REAL:: lambda
         real :: A
         REAL, INTENT(IN):: nom,lam
         INTEGER, INTENT(OUT):: C
         INTEGER:: i,B
         i = 0

          A = 0.0

          B=0

          DO WHILE (A<=nom)
          A = ((  (exp(-lam)) * (lam**i) ) /FACT(i)) + A
        B = B+1
        i = i+1
        C= B-1

         EnD DO

          END SUBROUTINE POISSONTRY




         INTEGER Function FACT(N)
         INTEGER, INTENT (IN):: N
         INTEGER:: J
         Fact = 1

         DO J = 0, N
         IF (J.NE.0) THEN

         Fact = Fact *J

         ELSE
         Fact = 1

         END IF

         END DO

         END FUNCTION FACT


 SUBROUTINE bitsub(variable)
 IMPLICIT NONE
```

```
INTEGER:: k, bits
INTEGER, INTENT(IN):: variable
Do k = 0,31
bits = btest1(variable, k)

!WRITE(unit = *, fmt = "(I)", advance = 'no') bits
END DO
END Subroutine bitsub


integer function tractlength(ve)
real::ve
integer:: leng, lambda
lambda = 352            !!!!!!!!!!!!!!parameter for tract length
tractlength = (-lambda*log(1-ve))
END FUNCTION tractlength



        SUBROUTINE fixednonsyn(n, mZyg, mGam, mChr,se)
    INTEGER, INTENT(IN):: mZyg, mGam, mChr
     INTEGER:: jChr,  jZyg, jGam,  i, first_test, start, chromosome
     INTEGER, DIMENSION(mZyg, mGam, mChr),  INTENT(IN) ::n
     REAL:: prop1, prop0,propsegr, nosites1, nosites0,no_fixed0, no_fixed1
     INTEGeR::  BB,AA,  bits1, bits2, ipos,  mult
      REAL::CC,F
     REAL, INTENT (IN), DIMENSION (0:31,mChr):: se
     REAL:: nuc
    INTEGER::      test, no_segregating, ende
     REAL :: theta
     REAL:: a,b,nucleotidediv
     REAL:: a1,a2,b1,b2,c1,c2,e1,e2, S, D, a11, btes, n_2, t, t1, t2, var,
        REAL:: sum_coeff, count_sum, average_sel_coeff

     sum_coeff = 0.0
      count_sum = 0.0

     no_fixed0 = 0.0
     no_fixed1 = 0.0
  CC = 0.0

 start = -1
     DO Chromosome = 1,3
     start = start + 1
Chr:    DO jChr = chromosome ,mChr, 3

posi:    Do i = start, 31, 3

        nosites0=0.0
        nosites1 = 0.0


        !! DIVERSITY PART:::::::::::::
         BB = 0

        Zygote:        DO jZyg = 1, 100
!PRINT*, "zyg", jZyg
        AA = 0

        bits1 = btest1(n(jZyg,1,jChr), i)


        AA = bits1
       ! PRINT*, "A", A


        BB = BB+AA              !sum of all "1" at that position

        mult = (100.0-BB)*BB  !no of pairwise diff. at that pos.

        END DO Zygote
         CC = CC + mult

         !!!!!END DIVERSITY





     first_test = btest1(n(1,1,jChr),i) !either 1 or 0
 IF (first_test==0) THEN

Zygotel:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)

     IF (test==first_test) THEN
     nosites0=nosites0+1.0
     ELSE
     exit Zygotel
     ENDIF

     END DO
     END DO Zygotel

     IF (nosites0 == (100.0)) THEN

     no_fixed0 = no_fixed0+1.0
     ENDIF

ELSEIF (first_test == 1) THEN
Zygotes:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)

     IF (test==first_test) THEN
     nosites1=nosites1+1.0
     ELSE
     exit Zygotes
     ENDIF

     END DO
     END DO Zygotes

     IF (nosites1 == (100.0)) THEN
     !PRINT*, s(i, jChr)
      sum_coeff = sum_coeff+ se(i,jChr)
        count_sum = count_sum + 1.0

     no_fixed1 = no_fixed1+1.0
     ENDIF

ENDIF

     END DO posi

     !FOR SECOND NON_SYN POSITION


     pos:    Do i = start+1, 31, 3

        nosites0=0.0
        nosites1 = 0.0

         !!!DIVERsiTY!!!!!!!!!!!!!!!!!!!!
       BB = 0

Zygotem:        DO jZyg = 1, 100
        AA = 0

        bits1 = btest1(n(jZyg,1,jChr), i)


        AA = bits1
       ! PRINT*, "A", A
```

```
        BB = BB+AA              !sum of all "1" at that position

        mult = (100.0-BB)*BB  !no of pairwise diff. at that pos.
             END DO Zygotem

      !  PRINT*, "B", B

        CC = CC + mult
!!!END DIVERITY!!!!!!!!!!!!!!!!!!!


     first_test = btest1(n(1,1,jChr),i) !either 1 or 0
  IF (first_test==0) THEN

Zy:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)

     IF (test==first_test) THEN
     nosites0=nosites0+1.0
     ELSE
     exit Zy
     ENDIF

     END DO
     END DO Zy

     IF (nosites0 == (100.0)) THEN

     no_fixed0 = no_fixed0+1.0
     ENDIF

ELSEIF (first_test == 1) THEN
Zygo:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)

     IF (test==first_test) THEN
     nosites1=nosites1+1.0
     ELSE
     exit Zygo
     ENDIF

     END DO
     END DO Zygo

     IF (nosites1 == (100.0)) THEN
  sum_coeff = sum_coeff+se(i,jChr)
      count_sum = count_sum+1.0
      no_fixed1 = no_fixed1+1.0
     ENDIF

ENDIF

     END DO pos

    END DO Chr
    END DO


!     FOR third chromosome, position 0
Ch:    DO jChr = 3 ,mChr, 3

 i = 0

  !!!DIVERsiTY!!!!!!!!!!!!!!!!!!!
       BB = 0

Zygoter:        DO jZyg = 1, 100
!PRINT*, "zyg", jZyg
        AA = 0

        bits1 = btest1(n(jZyg,1,jChr), i)


        AA = bits1

        BB = BB+AA              !sum of all "1" at that position

        mult = (100.0-BB)*BB  !no of pairwise diff. at that pos.
           END DO Zygoter

          CC = CC + mult
!!!END DIVERITY!!!!!!!!!!!!!!!!!!!





        nosites0=0.0
        nosites1 = 0.0

     first_test = btest1(n(1,1,jChr),i) !either 1 or 0
  IF (first_test==0) THEN

Zygot:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)
     IF (test==first_test) THEN
     nosites0=nosites0+1.0
     ELSE
     exit Zygot
     ENDIF

     END DO
     END DO Zygot

     IF (nosites0 == (100.0)) THEN
     no_fixed0 = no_fixed0+1.0
     ENDIF

ELSEIF (first_test == 1) THEN
Zygotess:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)

     IF (test==first_test) THEN
     nosites1=nosites1+1.0
     ELSE
     exit Zygotess
     ENDIF

     END DO
     END DO Zygotess

     IF (nosites1 == (100.0)) THEN
       sum_coeff = sum_coeff + se(i,jChr)
     count_sum = count_sum+1.0
     no_fixed1 = no_fixed1+1.0
     ENDIF

ENDIF

 END DO Ch

  F = 100.0*(100.0 - 1.0)*0.5
 nuc = CC/F

average_sel_coeff= sum_coeff/count_sum
PRINT*, " "
PRINT*, "average sel. coeff at fixed nonsyn.sites"
PRINT*, average_sel_coeff

 IF (mod (mChr, 3) == 0) THEN
prop0 = no_fixed0/((mChr*32)/1.5)
prop1 = no_fixed1/((mChr*32)/1.5)
ELSEIF (mod (mChr,3) ==2) THEN !mCHr = 2,5 etc. =>
```

```fortran
prop0 = no_fixed0/(   ( ( (mChr-2)*32) /1.5)+43)
prop1 = no_fixed1/(   ( ( (mChr-2)*32) /1.5)+43)
ELSEIF (mod (mChr,3) ==1) THEn ! mChr = 1,4 etc
prop0 = no_fixed0/(   (mChr - 1)*32/1.5 + 22)
prop1 = no_fixed1/(   (mChr - 1)*32/1.5 + 22)
ENDIF


propsegr = 1-prop0-prop1

IF (mod (mChr, 3) == 0) THEN
no_segregating = (  (32/1.5)*mChr)-no_fixed0-no_fixed1
ELSEIF (mod (mChr,3) ==2) THEN !mCHr = 2,5 etc. =>
no_segregating = (  (32/1.5) * (mChr -2)+ 43)-no_fixed0-no_fixed1
ELSEIF (mod (mChr,3) ==1) THEn ! mChr = 1,4 etc
no_segregating = (   (32/1.5 ) * (mChr -1) + 22)-no_fixed0-no_fixed1
ENDIF

     nucleotidediv=nuc/(32*mChr*0.6666666)



     PRINT*, "no. fixed nonsyn. sites", " for 0="
     PRINT*, no_fixed0
     PRINT*,  " and for 1="
     PRINT*,  no_fixed1
     PRINT*, "proportion of nonsyn. sites fixed:", "for 0"
     WRITE(unit = *, fmt = "(2F6.2)") prop0
     PRINT*, "and for 1"
     WRITE(unit = *, fmt = "(2F6.2)") prop1
     PRINT*, "proportion of nonsyn. sites that are segreagating:"
     WRITE(unit = *, fmt = "(2F6.2)")  propsegr
     PRINT*, "no_segregating"
     PRINT*,  no_segregating
     PRINT*, " "
     PRINT*, "nucleotidediv nonsyn"
     PRINT*,  nucleotidediv

     !!!TAJIMA:
      a1 = 0
      ende = (100.0)-1.0
        DO i = 1, ende
      a = (1.0/i)
      a1 = a1+a
      END DO

     theta = no_segregating/a1      !!!!!theta = S/(1/i)
     S = theta*a1
```

```
      a2 = 0

      DO i = 1, ende
      a11 = 1.0/(i**2)
      a2 = a11+a2
      END DO
       n_2= 100


      b1 = (n_2 + 1.0)/(3.0 * (n_2-1.0) )

      b2 = 2.0*(n_2*n_2 + n_2 + 3.0)/(9.0*n_2*(n_2-1.0))

      c1 = b1 - 1/a1

      c2 = b2 - ((n_2 + 2.0)/(a1*n_2)) + a2/(a1*a1)

      e1 = c1/a1

      e2 = c2/((a1*a1) + a2)

         var = SQRT((e1*S) +e2*S*(S-1.0))

      diff =  nuc -theta

      D= diff/var
     PRINT*, " "
     PRINT*, "Tajima's D nonsyn 10%"
      PRINT*, D
        k_min = no_segregating* (100.0-1.0)/F     != 2S/n    = S*(n-s)//n*(n

       D_min = abs((  k_min-(S/a1)) / var )
      D_rel = D/D_min
       PRINT*,"D_rel nonsynonymous"
       PRINT*,  D_rel
       D_rel = (nuc -theta) / abs(k_min-theta)
    !PRINT*,"D_rel", D_rel


   END SUBROUTINE fixednonsyn


  !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
  ! !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!



          SUBROUTINE fixedsyn(n, mZyg, mGam, mChr)
      INTEGER, INTENT(IN):: mZyg, mGam, mChr
```

```fortran
      INTEGER:: jChr,  jZyg, jGam,  i, first_test, start, chromosome
      INTEGER, DIMENSION(mZyg, mGam, mChr),  INTENT(IN) ::n
      REAL:: prop1, prop0,propsegr_syn, nositessyn1, nositessyn0,no_fixedsy
      INTEGeR::  BB,AA,  bits1, bits2, ipos,  mult
       REAL::CC,F

      REAL:: nuc
     INTEGER::      test, no_segregating, ende
      REAL :: theta
      REAL:: a,b,nucleotidediv
      REAL:: a1,a2,b1,b2,c1,c2,e1,e2, S, D, a11, btes, n_2, t, t1, t2, var,





      no_fixedsyn0 = 0.0
         no_fixedsyn1 = 0.0
           CC = 0
         start = 4
      DO chromosome = 1,2
      start = start - 2
      DO jChr = chromosome,mChr,3
      Do i = start,31,3

           !!!DIVERsiTY!!!!!!!!!!!!!!!!!!!!
        BB = 0

 Zygote:        DO jZyg = 1, 100
         AA = 0

         bits1 = btest1(n(jZyg,1,jChr), i)


         AA = bits1

         BB = BB+AA              !sum of all "1" at that position

         mult = (100-BB)*BB  !no of pairwise diff. at that pos.
              END DO Zygote

         CC = CC + mult
 !!!END DIVERITY!!!!!!!!!!!!!!!!!!!!


         nositessyn0=0.0
         nositessyn1 = 0.0

      first_test = btest1(n(1,1,jChr),i) !either 1 or 0
```

```
IF (first_test==0) THEN

Z:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)

     IF (test==first_test) THEN
     nositessyn0=nositessyn0+1.0
     ELSE
     exit Z
     ENDIF

     END DO
     END DO Z

     IF (nositessyn0 == (100)) THEN
     no_fixedsyn0 = no_fixedsyn0+1.0
     ENDIF

ELSEIF (first_test == 1) THEN
Zs:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)

     IF (test==first_test) THEN
     nositessyn1=nositessyn1+1.0
     ELSE
     exit Zs
     ENDIF

     END DO
     END DO Zs

     IF (nositessyn1 == (100)) THEN
       no_fixedsyn1 = no_fixedsyn1+1.0
     ENDIF

ENDIF

     END DO
        END DO
        END Do

!third chromsome


     DO jChr = 3,mChr,3
     Do i = 1,31,3
         !!!DIVERsiTY!!!!!!!!!!!!!!!!!!!
       BB = 0

Zygoten:        DO jZyg = 1, 100
        AA = 0

        bits1 = btest1(n(jZyg,1,jChr), i)


        AA = bits1

        BB = BB+AA              !sum of all "1" at that position

        mult = (100-BB)*BB  !no of pairwise diff. at that pos.
           END DO Zygoten

        CC = CC + mult
!!!END DIVERITY!!!!!!!!!!!!!!!!!!!


        nositessyn0=0.0
        nositessyn1 = 0.0

     first_test = btest1(n(1,1,jChr),i) !either 1 or 0
  !  PRINT*, first_test
IF (first_test==0) THEN

Za:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)

     IF (test==first_test) THEN
     nositessyn0=nositessyn0+1.0
     ELSE
     exit Za
     ENDIF

     END DO
     END DO Za

     IF (nositessyn0 == (100)) THEN
       no_fixedsyn0 = no_fixedsyn0+1.0
     ENDIF

ELSEIF (first_test == 1) THEN
Zas:     DO jZyg = 1,100
     DO jGam = 1,mGam
     test= btest1(n(jZyg, jGam, jChr),i)

     IF (test==first_test) THEN
```

```
      nositessyn1=nositessyn1+1.0
      ELSE
      exit Zas
      ENDIF

      END DO
      END DO Zas

      IF (nositessyn1 == (100)) THEN
      no_fixedsyn1 = no_fixedsyn1+1.0
      ENDIF

 ENDIF

      END DO
         END DO

  F = 100*(100 - 1.0)*0.5    !!!!!!! = n(n-1)/2

         nuc = CC/F

 if( mod( mChr, 3) == 0) THEN
 prop0 = no_fixedsyn0/((mChr*32)/3.0)
 prop1 = no_fixedsyn1/((mChr*32)/3.0)
  ELSEIF (mod (mChr,3) ==2) THEN !mCHr = 2,5 etc. =>
  prop0 = no_fixedsyn0/(  ((mChr-2)*32)/3.0 +21)
  prop1 = no_fixedsyn1/(  ((mChr-2)*32)/3.0 +21)
  ELSEIF (mod (mChr,3) ==1) THEN !mCHr = 1,4 etc.
  prop0 = no_fixedsyn0/(  ((mChr-1)*32)/3.0 + 10)
  prop1 = no_fixedsyn1/(  ((mChr-1)*32)/3.0 + 10)
  ENDIF

 propsegr_syn=1.0-prop0-prop1

 if( mod( mChr, 3) == 0) THEN
 no_segregating = ((32/3.0)*mChr)-no_fixedsyn0-no_fixedsyn1
  ELSEIF (mod (mChr,3) ==2) THEN !mCHr = 2,5 etc. =>
  no_segregating = (((32/3.0 )*(mChr-2))+21) -no_fixedsyn0-no_fixedsyn1
 ELSEIF (mod (mChr,3) ==1) THEN !mCHr = 1,4 etc.
 no_segregating = (((32/3.0 )*(mChr-1))+10) -no_fixedsyn0-no_fixedsyn1
  ENDIF

 nucleotidediv=nuc/(mChr*32*0.333333333)
 PRINT*, "no fixed synonymous sites for  0="
 PRINT*, no_fixedsyn0
 PRINT*,  " and  for 1="
  PRINT*,  no_fixedsyn1
    PRINT*, " "
  PRINT*, "proportion of sites fixed:", "for 0"
  WRITE(unit = *, fmt = "(2F6.2)") prop0
   PRINT*,  "and for 1"
   WRITE(unit = *, fmt = "(2F6.2)") prop1
   PRINT*, " "
  PRINT*, "proportion of syn. sites that are segregating:"
  WRITE(unit = *, fmt = "(2F6.2)") propsegr_syn
  PRINT*, "no segregating syn"
   PRINT*, no_segregating
   PRINT*, " "
  PRINT*," nucleotidediv syn"
  PRINT*,  nucleotidediv
  PRINT*, " "
  PRINT*, " "


 ! TAJIMA::

   a1 = 0
       ende = (100)-1.0
           DO i = 1, ende
       a = (1.0/i)
       a1 = a1+a
       END DO
          theta = no_segregating/a1      !!!!!theta = S/(1/i)
      S = theta*a1



      a2 = 0

      DO i = 1, ende
      a11 = 1.0/(i**2)
      a2 = a11+a2
      END DO
       n_2= 100


      b1 = (n_2 + 1.0)/(3.0 * (n_2-1.0) )

      b2 = 2.0*(n_2*n_2 + n_2 + 3.0)/(9.0*n_2*(n_2-1.0))

      c1 = b1 - 1/a1

      c2 = b2 - ((n_2 + 2.0)/(a1*n_2)) + a2/(a1*a1)

      e1 = c1/a1

      e2 = c2/((a1*a1) + a2)
      var = SQRT((e1*S) +e2*S*(S-1.0))
         diff =  nuc -theta

      D= diff/var
     PRINT*, " "
     PRINT*, " Tajima's D synonymous 10%"
      PRINT*, D

    k_min = no_segregating* (100.0-1.0)/F        != 2S/n   = S*(n-s)//n*(n
     D_min = abs((  k_min-(S/a1)) / var )
     D_rel = D/D_min
       PRINT*,"D_rel synonymous"
       PRINT*,  D_rel
     D_rel = (nuc -theta) / abs(k_min-theta)
   !  PRINT*,"D_rel", D_rel



   END SUBROUTINE fixedsyn


   !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

   !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
   !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

  SUbroutine LD_selected(n,mZyg,mGam,mChr)
 INTEGER, INTENT(IN):: mZyg, mGam, mChr
      INTEGER:: jChr,  jZyg, jGam, jpos
      INTEGER, DIMENSION(mZyg, mGam, mChr),  INTENT(IN) ::n
 real:: x11,x10,x01,x00,D, p0, p1, Dprime, Dmax,q0,q1, Dsum, Daverage
 INTEGER:: test, test2, cou, sites

 Dsum = 0
 cou=0

 DO jChr = 1,mChr,3
 DO jpos = 0,30,3

 cou=cou+1
 x11 = 0.0
 x10 = 0.0
 x01 = 0.0
 x00 = 0.0

 sites=0

  Do jZyg = 1,mZyg
 Do jGam = 1,mGam
```

```fortran
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr),jpos+1)

IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO

x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)

!allele frequencies:
p0 = x01+x00
p1 = x11+x10
q0 = x10 + x00
q1 = x11+x01

!CALCULATE D

D = x11-(p1*q1)

!CALCULATE Dmax
IF (D .ge. 0) THEN

IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
Dmax = p1*q1
ENDIF

ENDIF
```

```
!CALCULATE Dprime
Dprime= D/Dmax

IF (Dmax==0) THEN

cou=cou-1
Dprime=0
ENDIF


Dsum = Dsum + Dprime

END DO
END DO

!PRINT*, "CHROMOSOME2"

DO jChr = 1,mChr,3
DO jpos = 1,28,3

cou=cou+1
x11 = 0.0
x10 = 0.0
x01 = 0.0
x00 = 0.0

sites=0

 Do jZyg = 1,mZyg
Do jGam = 1,mGam
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr),jpos+1)

IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO
```

```fortran
!haplotype frequencies:

x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)

!allele frequencies:
p0 = x01+x00
p1 = x11+x10
q0 = x10 + x00
q1 = x11+x01

!CALCULATE D

D = x11-(p1*q1)

!CALCULATE Dmax
IF (D .ge. 0) THEN

IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
Dmax = p1*q1
ENDIF

ENDIF


!CALCULATE Dprime
Dprime= D/Dmax


IF (Dmax==0) THEN

cou=cou-1
Dprime=0

ENDIF


Dsum = Dsum + Dprime
```

```
END DO
END DO

!!!!!!!!!!!!!!!!!!!!FOR position 31 on second chromosome etc.

IF (mCHR.ge.3) THEN
DO jChr = 1,mChr-1,3
 jpos = 31

cou=cou+1
x11 = 0.0
x10 = 0.0
x01 = 0.0
x00 = 0.0

sites=0

 Do jZyg = 1,mZyg
Do jGam = 1,mGam
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr+1),0)

IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO

!haplotype frequencies:
x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)

!allele frequencies:
p0 = x01+x00
p1 = x11+x10
q0 = x10 + x00
q1 = x11+x01
```

```
!CALCULATE D

D = x11-(p1*q1)

!CALCULATE Dmax
IF (D .ge. 0) THEN

IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
Dmax = p1*q1
ENDIF

ENDIF

!CALCULATE Dprime
Dprime= D/Dmax

IF (Dmax==0) THEN
cou=cou-1
Dprime=0

ENDIF


Dsum = Dsum + Dprime

END DO

ENDIF

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
DO jChr = 3,mChr,3
 DO jpos = 2,29,3

cou=cou+1
x11 = 0.0
x10 = 0.0
x01 = 0.0
x00 = 0.0

sites=0
```

```
 Do jZyg = 1,mZyg
Do jGam = 1,mGam
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr),jpos+1)

IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO


x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)

!allele frequencies:
p0 = x01+x00
p1 = x11+x10
q0 = x10 + x00
q1 = x11+x01

!CALCULATE D

D = x11-(p1*q1)

!CALCULATE Dmax
IF (D .ge. 0) THEN

IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
```

```
Dmax = p1*q1
ENDIF

ENDIF


!CALCULATE Dprime
Dprime= D/Dmax


IF (Dmax==0) THEN
cou=cou-1
Dprime=0
ENDIF


Dsum = Dsum + Dprime
END DO
END DO
!PRINT*, " "

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
PRiNT*, "no. of comparisons selected"
PRINT*, cou
Daverage = Dsum/cou
PRINT*, "Daverage selected"
PRINT*, Daverage



END SUBROUTINE LD_selected

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!



 SUbroutine LD_neutral(n,mZyg,mGam,mChr)
INTEGER, INTENT(IN):: mZyg, mGam, mChr
     INTEGER:: jChr,  jZyg, jGam, jpos
     INTEGER, DIMENSION(mZyg, mGam, mChr),  INTENT(IN) ::n
real:: x11,x10,x01,x00,D, p0, p1, Dprime, Dmax,q0,q1, Dsum, Daverage
INTEGER:: test, test2, cou, sites

Dsum = 0
cou=0

DO jChr = 1,mChr,3
DO jpos = 2,26,3
```

```fortran
cou=cou+1
x11 = 0.0
x10 = 0.0
x01 = 0.0
x00 = 0.0

sites=0

 Do jZyg = 1,mZyg
Do jGam = 1,mGam
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr),jpos+3)

IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO

!PRINT*, " "

x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)

!allele frequencies:
p0 = x01+x00
p1 = x11+x10
q0 = x10 + x00
q1 = x11+x01


!CALCULATE D

D = x11-(p1*q1)

!CALCULATE Dmax
IF (D .ge. 0) THEN
```

```fortran
IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
Dmax = p1*q1
ENDIF

ENDIF

!CALCULATE Dprime
Dprime= D/Dmax



IF (Dmax==0) THEN
cou=cou-1
Dprime=0
ENDIF


Dsum = Dsum + Dprime
END DO
END DO
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
IF ( mCHr .ge. 2) THEN
!FOR POSITION 29 on Chromosome 1 etx
DO jChr = 1,mChr-1,3
 jpos = 29

cou=cou+1
x11 = 0.0
x10 = 0.0
x01 = 0.0
x00 = 0.0

sites=0

 Do jZyg = 1,mZyg
Do jGam = 1,mGam
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr+1),0)
```

```
IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO


x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)

!allele frequencies:
p0 = x01+x00
p1 = x11+x10
q0 = x10 + x00
q1 = x11+x01

!CALCULATE D

D = x11-(p1*q1)
!CALCULATE Dmax
IF (D .ge. 0) THEN

IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
Dmax = p1*q1
ENDIF

ENDIF

!CALCULATE Dprime
Dprime= D/Dmax
```

```
IF (Dmax==0) THEN
cou=cou-1
Dprime=0
ENDIF


Dsum = Dsum + Dprime
END DO
ENDIF
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
DO jChr = 2,mChr,3
DO jpos = 0,27,3

cou=cou+1
x11 = 0.0
x10 = 0.0
x01 = 0.0
x00 = 0.0

sites=0

 Do jZyg = 1,mZyg
Do jGam = 1,mGam
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr),jpos+3)

IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO

!PRINT*, " "

x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)
!allele frequencies:
p0 = x01+x00
```

```
p1 = x11+x10
q0 = x10 + x00
q1 = x11+x01


!CALCULATE D

D = x11-(p1*q1)

!CALCULATE Dmax
IF (D .ge. 0) THEN

IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
Dmax = p1*q1
ENDIF

ENDIF


!CALCULATE Dprime
Dprime= D/Dmax


IF (Dmax==0) THEN
cou=cou-1
Dprime=0
ENDIF


Dsum = Dsum + Dprime
END DO
END DO
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!FOR POSITION 30 on chr 2 etc
IF (mChr .ge. 3) THEN
DO jChr = 2,mChr-1,3
 jpos = 30

cou=cou+1
x11 = 0.0
```

```
x10 = 0.0
x01 = 0.0
x00 = 0.0

sites=0

 Do jZyg = 1,mZyg
Do jGam = 1,mGam
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr+1),1)

IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO


x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)


!allele frequencies:
p0 = x01+x00
p1 = x11+x10
q0 = x10 + x00
q1 = x11+x01


!CALCULATE D

D = x11-(p1*q1)

!CALCULATE Dmax
IF (D .ge. 0) THEN

IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
```

```
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
Dmax = p1*q1
ENDIF

ENDIF


!CALCULATE Dprime
Dprime= D/Dmax


IF (Dmax==0) THEN
cou=cou-1
Dprime=0
ENDIF


Dsum = Dsum + Dprime
END DO
ENDIF
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
DO jChr = 3,mChr,3
DO jpos = 1,28,3

cou=cou+1
x11 = 0.0
x10 = 0.0
x01 = 0.0
x00 = 0.0

sites=0

 Do jZyg = 1,mZyg
Do jGam = 1,mGam
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr),jpos+3)

IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
```

```
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO

x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)
!allele frequencies:
p0 = x01+x00
p1 = x11+x10
q0 = x10 + x00
q1 = x11+x01

!CALCULATE D

D = x11-(p1*q1)


!CALCULATE Dmax
IF (D .ge. 0) THEN

IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
Dmax = p1*q1
ENDIF

ENDIF



!CALCULATE Dprime
Dprime= D/Dmax


IF (Dmax==0) THEN
cou=cou-1
```

```
Dprime=0
ENDIF


Dsum = Dsum + Dprime
END DO
END DO
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
IF (mChr .ge. 4) THEN
!FOR POSITION 31 on chr 3 etc
DO jChr = 3,mChr-1,3
 jpos = 31

cou=cou+1
x11 = 0.0
x10 = 0.0
x01 = 0.0
x00 = 0.0

sites=0

 Do jZyg = 1,mZyg
Do jGam = 1,mGam
sites = sites +1
test = btest1(n(jZyg, jGam,jChr),jpos)
test2=btest1(n(jZyg, jGam,jChr+1),2)

IF ((test ==1) .and. (test2==1)) THEN
x11 = x11+1.0
ELSEiF ((test ==1) .and. (test2==0)) THEN
x10 = x10+1.0
ELSEiF ((test ==0) .and. (test2==1)) THEN
x01 = x01+1.0
ELSEiF ((test ==0) .and. (test2==0)) THEN
x00 = x00+1.0

ENDIF
END DO
END DO


x11 = x11/(sites)
x10= x10/(sites)
x01 = x01/(sites)
x00 = x00/(sites)

!allele frequencies:
p0 = x01+x00
p1 = x11+x10
```

```
q0 = x10 + x00
q1 = x11+x01


!CALCULATE D

D = x11-(p1*q1)
!CALCULATE Dmax
IF (D .ge. 0) THEN

IF( (p0*q1) .le. (p1*q0) )THEN
Dmax = p0*q1
ELSE
Dmax = p1*q0
ENDIF

ELSEIF (D .lt. 0) THEN
IF( (p0*q0) .le. (p1*q1) )THEN
Dmax = p0*q0
ELSE
Dmax = p1*q1
ENDIF

ENDIF


!CALCULATE Dprime
Dprime= D/Dmax


IF (Dmax==0) THEN
cou=cou-1
Dprime=0
ENDIF


Dsum = Dsum + Dprime
!PRINT*, "cou", cou
END DO
ENDIF
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

PRiNT*, "no.of comparisons neutral"
PRINT*, cou
Daverage = Dsum/cou
PRINT*, "Daverage neutral"
PRINT*, Daverage
```

```
END SUBROUTINE LD_neutral


SUBROUTINE Bsel(n, mZyg, mGam, mChr, sel, recf, GCf)
INTEGER, INTENT(IN) :: mChr, mZyg, mGam
REAL, INTENt (IN) :: recf, GCf
INTEGER, DIMENSION(mZyg, mGam, mChr),  INTENT(IN) ::n
REAL, dimension(0:31,mCHR), INTENT(IN):: sel
INTEGER::  multiplier, jChr, iPos, jPos, start, chromosome
INTEGER:: bin
INTEGER:: counts
REAL:: exponentsum, Z, r,u, s
REAL, DIMENSIOn (0:9):: B, nucleotidediv, rel_red,compare_B
INTEGER:: start_bin, ende_bin,BB, mult, bits1,bits2, bin_size, jZyg
     INTEGER:: AA, F, rounds
     REAL:: nuc,div_neutral_expected,CC,distance

counts = 0


!10 data points to calculate B

PRINT*, "expected B"
multi: DO multiplier = 0,9
counts = 0
bin = multiplier
jChr = (mChr/10-(mChr/20))+(mChr/10)*multiplier ! Chromosome on Which foca

 IF (mod (jChr, 3) == 0) THEN
 jPos = 16
 ELSEIF (mod (jChr,3) ==2) THEN !mCHr = 2,5 etc. =>
 jPos = 15
ELSEIF (mod (jChr,3) ==1) THEn ! mChr = 1,4 etc
jPos = 17
ENDIF


Exponentsum = 0.0

start = -1
     DO Chromosome = 1,3
     start = start + 1
Chr:    DO iChr = chromosome ,mChr, 3

posi:    Do i = start, 31, 3!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```

```
!Calculate exponent:

distance = abs(jChr*32-iChr*32+jPos-i)
counts = counts+1
Z = distance * recf + 2* (352* GCf) *(  1 - exp(-(distance/352.0))  ) !inc
!Z = distance *recf   !=distance times rec rate

r = 0.5* (1-exp(-(2*z) ) )
u = 0.0000104
s = sel(i, iChr)
exponentsum = exponentsum+ u/( s* (  ( 1+ (r*(1-s)/s)  )**2) )
IF (s==0 ) THEn
!PRINT*, "s alamr!"
ENDIF

ENDDO posi

!FOR SECOND NON_SYN POSITION
 pos:    Do i = start+1, 31, 3

 distance = abs(jChr*32-iChr*32+jPos-i)

distance = abs(jChr*32-iChr*32+jPos-i)
counts = counts+1
Z = distance * recf + 2* (352* GCf) *(  1 - exp(-(distance/352.0))  ) !inc
!Z = distance *recf   !=distance times rec rate

r = 0.5* (1-exp(-(2*z) ) )
u = 0.0000104
s = sel(i, iChr)
exponentsum = exponentsum+ u/( s* (  ( 1+ (r*(1-s)/s)  )**2) )
IF (s==0 ) THEn
!PRINT*, "s alamr!"
ENDIF
!PRINT*, "exponentsum", exponentsum


 ENDDO Pos
 ENDDO Chr

  ENDDO

 !      FOR third chromosome, position 0
Ch:    DO iChr = 3 ,mChr, 3
 i = 0
 !PRINT*, jChr, jPos
 !PRINT*, iChr, i
distance = abs(jChr*32-iChr*32+jPos-i)
```

```
counts = counts+1
Z = distance * recf + 2* (352* GCf) *(  1 - exp(-(distance/352.0))  ) !inc
!Z = distance *recf   !=distance times rec rate
r = 0.5* (1-exp(-(2*z) ) )

u = 0.0000104

s = sel(0, iChr)
IF (s==0 ) THEn
!PRINT*, "s alamr!"
ENDIF
!PRINT*, "exponentsum before summations", exponentsum
exponentsum = exponentsum+ u/( s* (  ( 1+ (r*(1-s)/s)  )**2) )


 END DO Ch


 PRINT*, "exponentsum", exponentsum
 B(bin) = exp(-exponentsum)
! PRINT*, " expected B BIN", multiplier, "="
 PRINT*, B(bin)
! PRINT*, "count", counts
 ENDDO multi

     !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!


  !CaLCULATE OBSERVED sysnonymous  DIVERSITY in BINS


     start_bin = 1
     ende_bin = mChr/10
     bin = 0
     bin_size = mChr/10

   !  PRINT*, BIN_SIZE, "BIN_SIZE"

bins:     Do rounds = 0,9
CC=0
Chr2:     DO jChr = start_bin, ende_bin
   !  PRINT*, " "
   !  PRINT*, "jChr", jChr
   !  PRINT*, " "

    !!!!!!!!!!calc DIV

     IF (mod (jChr,3) ==0 )THEN   ! Chr 3 etc
     DO i = 1,31,3
```

```
      IF (sel(i,jChr) .ne. 0) THEN
      PRINT*, "ALEMR"
      ENDIF

     ! PRINT*, i
         !! DIVERSITY PART:::::::::::::
           BB = 0

       DO jZyg = 1, mZyg
         AA = 0

         bits1 = btest1(n(jZyg,1,jChr), i)

         AA = bits1
        ! PRINT*, "A", A


         BB = BB+AA              !sum of all "1" at that position

         mult = (mZyg-BB)*BB  !no of pairwise diff. at that pos.
       !  PRINT*, "mult", mult
        ! PRINT*, " "
         END DO

       !  PRINT*, "B", B

         CC = CC + mult

                  !!!!!END DIVERSITY
         ENDDO
         ELSEIF (mod (jChr,3) ==1) THEN   ! Chr 1 etc
      DO i = 2,29,3
      IF (sel(i,jChr) .ne. 0) THEN
      PRINT*, "ALEMR"
      ENDIF

    !  PRINT*, i
         !! DIVERSITY PART:::::::::::::
         BB = 0

       DO jZyg = 1, mZyg
         AA = 0

         bits1 = btest1(n(jZyg,1,jChr), i)


         AA = bits1
        ! PRINT*, "A", A
```

```
         BB = BB+AA              !sum of all "1" at that position

         mult = (mZyg-BB)*BB  !no of pairwise diff. at that pos.
       !  PRINT*, "mult", mult
        ! PRINT*, " "
         END DO

       !  PRINT*, "B", B

         CC = CC + mult

                  !!!!!END DIVERSITY
         ENDDO

         ELSEIF (mod (jChr,3) ==2) THEN   ! Chr 2 etc
      DO i = 0,30,3
   !   PRINT*, i
      IF (sel(i,jChr) .ne. 0) THEN
      PRINT*, "ALEMR"
      ENDIF

         !! DIVERSITY PART:::::::::::::
         BB = 0

       DO jZyg = 1, mZyg
         AA = 0

         bits1 = btest1(n(jZyg,1,jChr), i)


         AA = bits1
        ! PRINT*, "A", A


         BB = BB+AA              !sum of all "1" at that position

         mult = (mZyg-BB)*BB  !no of pairwise diff. at that pos.
       !  PRINT*, "mult", mult
        ! PRINT*, " "
         END DO

       !  PRINT*, "B", B

         CC =  CC + mult

                  !!!!!END DIVERSITY
         ENDDO
```

```fortran
        ENDIF


         F = mZyg*(mZyg - 1.0)*0.5    !!!!!!! = n(n-1)/2

        nuc = CC/F
       !  PRINT*, F, CC, nuc
        nucleotidediv(bin)=nuc/(bin_size*32*0.333333333)
             ENDDO Chr2


       start_bin = start_bin + mChr/10
       ende_bin = ende_bin + mChr/10
     bin = bin +1

        ENDDO bins



    !    COMPARE exp and obs ntdiv


    div_neutral_expected = mZyg *2* 0.0000104   !expected Without BGS
    !PRINT*, div_neutral_expected

    PRINT*, "observed B "
    DO bin = 0,9
    rel_red (bin)= nucleotidediv(bin)/div_neutral_expected   !observed B
  !   PRINT*, "observed B ", "BIN", BIN
     PRINT*, rel_red (bin)
    ENDDO

     PRINT*, "observed red. in ntdiv/ expected B in bin"
    Do bin = 0,9
    compare_B(bin) = rel_red (bin)/B(bin)
  !  PRINT*, "observed red. in ntdiv/ expected B in bin", bin
    PRINT*, compare_B(bin)
    ENDDO




END SUBROUTINE Bsel

INTEGER FUNCTION Btest1(input, posit)
INTEGER, INTENT (IN) :: input, posit
LOGICAL:: testtt
testtt = btest(input, posit)
```

```
IF (testtt) then
btest1 = 1
ELSE
btest1 = 0
ENDIF

END FUNCTION Btest1



SUBROUTINE set_neutral(n,mZyg,mGam,mChr)
INTEGER, INTENT(IN):: mZyg, mGam, mChr
    INTEGER:: jChr,  jZyg, jGam, jpos
     INTEGER, DIMENSION(mZyg, mGam, mChr),  INTENT(INOUT) ::n
     REAL:: v,w,x,B,A,C,Y,D
     INTEGER:: st, chrom,i



i = 0
 st = 3
 Do chrom = 1,3,2
 st = st-1

Chr: DO jChr = chrom ,mChr,3
pos: Do jPos = st,31,3                  !only neutral sites though
!PRINT*,  jChr, jPos
Call Random_number(v)

Set1: IF (v .le. 0.4293757229) THEn     !do nothinbg => all remain zero

ElseIF ((v .gt. 0.4293757229 ) .and. ( v .le. 0.8587514458) ) THEn

DO jZyg = 1,mZyg
Do jGam = 1,mGam

n(jZyg,jGam,jChr) = ibset(n(jZyg,jGam,jChr), jPos)      !jPos is set to "o

END DO
END DO

ELSEIF (v .gt. 0.8587514458) THEN       !Polymorphic! decide how many "one'
!PRINT*, "polymorphic at pos", jPos
call random_number(w)

B = 1.0/(mZyg)
!PRINT*, "B", B
C = 0.0
A = 0.0
```

```
 D = 0.0
 !PRINT*, "w", w

 DO WHILE (C.le. w)
 !PRINT*,"C", C
 A =(1.0/B)/7484.47086   ! = SUM(i/2N)   ,i=1,2N-1 !!!!!!!!!!cvhange!!!!!!!
 C = A+C
 !C = Stammfunction
 D = B
 B = B+ (1/(mZyg))

 ENDDO
 !PRINT*, D
 !PRINT*, "Stammfunctionsvalue" ,C
 !PRINT*, "x-value D", D

 ! Now now the FRACTION  of gametes that carry "one" at jPos, C

 DO jZyg = 1,mZyg
 Do jGam = 1,mGam

 Call random_number(y)
 IF (y .le. D) THEN      !Pos set to "one"
 n(jZyg,jGam,jChr)= ibset(n(jZyg,jGam,jChr), jPos)       !jPos is set to "o
 ENDIF

 END DO
 END DO



 ENDIF SET1
 ENDDO POS
 ENDDo Chr
 ENDDO

 !FOR SECOND CHROMOSOME:
 Chr1: DO jChr = 2 ,mChr,3
 pos1: Do jPos = 0,31,3                  !only neutral sites though
 !PRINT*,jChr, jPos
 Call Random_number(v)

 Set2: IF (v .le. 0.429) THEn    !do nothing => all remain zero
 !PRINT*, "all zero at pos", jPos

 ElseIF  ((v .gt. 0.429 ) .and. ( v .le. 0.858) ) THEn
 !PRINT*, "all one at pos", jPos
 DO jZyg = 1,mZyg
 Do jGam = 1,mGam
```

```
n(jZyg,jGam,jChr) = ibset(n(jZyg,jGam,jChr), jPos)      !jPos is set to "o

END DO
END DO

ELSEIF (v .gt. 0.858) THEN      !Polymorphic! decide how many "one"
!PRINT*, "polymorphic at pos", jPos
call random_number(w)

B = 1.0/(mZyg)
!PRINT*, "B", B
C = 0.0
A = 0.0
D = 0.0
!PRINT*, "w", w

DO WHILE (C.le. w)
!PRINT*,"C", C
A =(1.0/B)/7484.47086   ! = SUM(i/2N)   ,i=1,2N-1 !!!!!!!!!cvhange!!!!!!
C = A+C
!C = Stammfunction
D = B
B = B+ (1/(mZyg))

ENDDO
!PRINT*, "Stammfunctionsvalue" ,C
!PRINT*, "x-value D", D
!PRINT*, D


! Now now the FRACTION  of gametes that carry "one" at jPos, C

DO jZyg = 1,mZyg
Do jGam = 1,mGam

Call random_number(y)
IF (y .le. D) THEN      !Pos set to "one"
n(jZyg,jGam,jChr)= ibset(n(jZyg,jGam,jChr), jPos)       !jPos is set to "o
ENDIF

END DO
END DO



ENDIF SET2
ENDDO POS1
ENDDo Chr1
```

```
 END SUBROUTINE set_neutral


 SUBROUTINE set_selected(n,se, mZyg,mGam,mChr)
 INTEGER, INTENT(IN):: mZyg, mGam, mChr
     INTEGER:: jChr,  jZyg, jGam, jpos
      INTEGER, DIMENSION(mZyg, mGam, mChr),  INTENT(INOUT) ::n
      REAL, DIMENSION(0:31, mChr), INTENT(IN) ::se
      REAL::q
      INTEGER:: chroma
     ! PRINT*," NONSYNONUMOUS"

         DO chroma = 1,3,2
         !PRINT*, "CHROMA", chroma
         CHr : Do jChr = chroma, mChr,3

         pos: Do jPos = 0,31,3
         !PRINT*, jChr, jPos
         IF (se(jPos,jChr) .ne. 0.0) THEN
         q = 0.0000104/(se(jPos,jChr))
         ELSE
         q = 0.0
         ENDIF

         !PRINT*, "q", q

         DO jZyg = 1,mZyg
         DO iGam = 1, mGam

         Call Random_number(v)

         IF (v .le. q) THEN
         n(jZyg,jGam,jChr) = ibset(n(jZyg,jGam,jChr), jPos)
         ENDIF

         ENDDO
         ENDDO

         ENDDO pos
         ENDDO Chr
         ENDDO

         !!!!!!!!!!!!!!!!!
         DO chroma = 1,2
         !PRINT*, "CHROMA", chroma
         CHr2 : Do jChr = chroma, mChr,3
```

```
pos2: Do jPos = 1,31,3
!PRINT*, jChr, jPos
IF (se(jPos,jChr) .ne. 0.0) THEN
q = 0.0000104/(se(jPos,jChr))
ELSE
q = 0.0
ENDIF

!PRINT*, "q", q

DO jZyg = 1,mZyg
DO iGam = 1, mGam

Call Random_number(v)

IF (v .le. q) THEN
n(jZyg,jGam,jChr) = ibset(n(jZyg,jGam,jChr), jPos)
ENDIF

ENDDO
ENDDO

ENDDO pos2
ENDDO Chr2
ENDDO
!!!!!!!!!!!!!!!!!!!!!
DO chroma = 2,3
!PRINT*, "CHROMA", chroma
CHr3 : Do jChr = chroma, mChr,3

pos3: Do jPos = 2,31,3
!PRINT*, jChr, jPos
IF (se(jPos,jChr) .ne. 0.0) THEN
q = 0.0000104/(se(jPos,jChr))
ELSE
q = 0.0
ENDIF

!PRINT*, "q", q

DO jZyg = 1,mZyg
DO iGam = 1, mGam

Call Random_number(v)

IF (v .le. q) THEN
n(jZyg,jGam,jChr) = ibset(n(jZyg,jGam,jChr), jPos)
ENDIF
```

```
ENDDO
ENDDO

ENDDO pos3
ENDDO Chr3
ENDDO


END SUBROUTINE set_selected




END PROGRAM projectrec2
```