The University Of Sheffield.

Quantifying the effects of natural selection and other evolutionary forces on the genome

By:

Benjamin C Jackson

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield

Faculty of Science

Department of Animal and Plant Sciences

11th July 2016

Acknowledgements

My thanks are due firstly to Kai Zeng, whose supervision of this PhD has been encouraging, helpful and patient. I've learnt a lot in the past three and a half years, in no small part due to Kai.

I would also like to thank Brian and Deborah Charlesworth, and their groups in Edinburgh, for data, help, discussion and feedback on all three of the main chapters below. In particular José Campos provided data and his bioinformatics expertise for Chapters 2 and 4, and much less of this thesis would have been possible if it wasn't for his input. Kang-Wook Kim kindly let me analyse his Gouldian Finch dataset in Chapter 3, and the *Drosophila Population Genomics Project* made the *D. melanogaster* dataset analysed in Chapter 2 publicly available.

Perhaps most importantly, I would like to thank my friends and colleagues in Sheffield, many of who have now left, and whom I miss. Often they helped me with science, sometimes they helped me to get through the working day. Mostly, they shared a lot of happy times with me inside and outside the department.

Maybe the single best thing about living in Sheffield for the duration of this PhD has been the Peak District National Park. Firstly, I'd like to thank Roger and Alison Butlin for trusting me at various times with their house, cat, and chickens, and so occasionally allowing me to live in the middle of it. Mainly, I'd like to acknowledge the soothing influence of its wilder inhabitants. Red Deer, Adders, Nightjars, Mountain Hares, Short-Eared Owls, Dippers, Hen Harriers, Pied Flycatchers, Redstarts, Cuckoos, and Woodcock are now species that are tied by happy memories to Sheffield. They, and the landscape they inhabit, improved my quality of life in an immeasurable way.

Abstract

Genomic or detailed genetic data are now increasingly available for both model and non-model organisms, and this is allowing researchers to investigate the role that different forces play in shaping the evolution of species' genomes. In this thesis I use large genetic and genomic datasets to attempt to quantify the action of natural selection and other forces on patterns of genetic polymorphism, differentiation and divergence in the natural world. These other forces include mutation, GC-biased gene conversion, genetic drift, and demography. Some of these forces act in the same direction in terms of their effect on the composition of genomes, and therefore the action of one may confound the detection of the action of others. To tell them apart, I often use a comparative method between different classes of sites, or between sites that differ in the extent to which they are subject to the relevant evolutionary pressures, as well as by using both polymorphism and divergence data. Where possible, I incorporate the effects of demography into these analyses. In the three main chapters that follow, I investigate 1) the forces that affect genetic differentiation between two populations of *Drosophila melanogaster*; 2) the forces underlying a naturally occurring colour polymorphism in the Australian finch *Erythrura gouldiae*; and 3) the forces which affect evolution at 4-fold degenerate sites in ancestral populations of *D. melanogaster* and *D. simulans*.

Contents

Statement of intellectual contribution

As expressed in the Acknowledgements, the work presented in this thesis owes much to the supervision of Kai Zeng and the help of several collaborators. The three main Chapters (2-4) presented below are written in the style of scientific papers, and the collaborating authors are listed at the beginning of each. The text of Chapter 2 is the same as in its published form (Heredity **114:**163-174), and the text of Chapter 4 should be submitted for publication largely as it stands. On top of supervision, the following specific contributions were made to these Chapters:

In Chapter 2, José Campos provided the coding sequence (CDS) and intronic alignments in FASTA format that were used for the subsequent analyses. Kai Zeng derived the dependence of estimates of F_{ST} on the minor allele frequency.

In Chapter 3, Kang-Wook Kim provided the alignments of the *Red* and reference loci in FASTA format that were used for the subsequent analyses. Kai Zeng derived the maximum likelihood estimates of allele frequencies; the maths underlying the estimate of θ incorporating a change in population size; and the maths underlying the variances and estimates of T and θ for a window of the *Red* locus, using the modified implementation of the HKA test that incorporates a change in population size, recombination, and non-random sampling.

In Chapter 4, José Campos converted multisample Variant Call Format (VCF) files into CDS and intronic alignments in FASTA format that were used for the subsequent analyses, using annotations from Flybase. Kai Zeng and Brian Charlesworth contributed the maths underlying the parts of the Discussion and Supplementary Text that pertains to distinguishing the effects of a reduction in the strength of selection from a change in mutational bias.

Chapter 1. General Introduction

In the late 1980s and early 1990s, a change in emphasis began in the field of molecular evolutionary biology, away from the neutralist perspective on the nature of the forces that determine molecular evolution, towards a selectionist one. This change came about partly as a result of the transition from allozymes to DNA-based genetic markers in population genetics studies, and the resulting finding of a positive association between the amount of genetic diversity and the level of genetic recombination in *Drosophila* (Aguade et al. 1989; Stephan and Langley 1989; Begun and Aquadro 1992). This pattern is expected if natural selection is pervasive in genomes (Hahn 2008).

Two central tenets of the neutral theory of molecular evolution (Kimura 1983) are that most polymorphism segregating within species has no effect on fitness, and that most differences between species have evolved as a result of genetic drift – i.e. natural selection does not play an important role in evolution (this is not the same as saying there is no selective constraint). Both of these predictions have also suffered in the past two decades. There is good evidence that the proportion of substitutions between species which are adaptive (α) is not negligible. In the human lineage, α has been estimated at between 10% and 40% (Fay et al. 2001; but see The Chimpanzee Sequencing and Analysis Consortium 2005; Boyko et al. 2008; Eyre-Walker and Keightley 2009). In the flowering plant *Capsella grandiflora*, α has been estimated at 40% (Slotte et al. 2010); in *Drosophila*, between 40-50% (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Andolfatto 2005; Welch 2006; Begun et al. 2007) and at more than 50% in *Mus musculus castaneus* (Halligan et al. 2010) and *Escherichia coli* (Charlesworth and Eyre-Walker 2006). There are also examples of balanced polymorphisms maintained by natural selection, for example in *Drosophila* (Kreitman and Hudson 1991); *Arabidopsis* (Bakker et al. 2006); *Caenorhabditis elegans* (Seidel et al. 2008) and domesticated sheep (Johnston et al. 2013), although these are specific examples, and whether balancing selection is actually widespread in nature is unclear.

As the move from allozymes to DNA markers provided impetus for the acceptance that natural selection is an importance force in molecular evolution, so has the availability of whole genome sequences cemented that view (Begun et al. 2007; Hahn 2008). Those first studies in *Drosophila* on the relationship between genetic diversity and

recombination rate, which used at most tens of genes (Aguade et al. 1989; Stephan and Langley 1989; Begun and Aquadro 1992), have been supplemented by work encompassing whole genomes, with thousands of coding sequence alignments and thousands or more single nucleotide polymorphisms (SNPs). The same positive relationship between diversity and recombination has been found in genomic studies of *Drosophila simulans* (Begun et al. 2007), *D. persimilis* (Stevison and Noor 2010), *Caenorhabditis briggsae* (Cutter and Choi 2010) and *Saccharomyces cerevisae* (Cutter and Moses 2011), amongst others (Cutter and Payseur 2013). In these studies, recombination was not correlated with divergence, so the best explanation for this pattern is that selection is common and that its effect on linked sites is modulated by recombination (Cutter and Payseur 2013).

In fact, different forms of natural selection can produce the pattern described above – both positive selection on favourable alleles (Maynard Smith and Haigh 1974) and negative selection against deleterious alleles (Charlesworth et al. 1993) reduce variation at linked sites, and so result in the same positive relationship between recombination and genetic diversity. Which force prevails is an open question, the answer to which probably varies taxon by taxon (Hahn 2008; Corbett-Detig et al. 2015). Hahn (2008) cited contrasting examples in humans, where a model of selection against deleterious mutations and the associated effect on linked sites (background selection; BGS (Charlesworth et al. 1993)) fitted polymorphism data (Reed et al. 2005; Lohmueller et al. 2011), compared to *Drosophila simulans* where recurrent bouts of positive selection fitted the data better (Begun et al. 2007; Sattath et al. 2011). More recently, Cutter and Choi (2010) found that the expectations under BGS are sufficient to explain patterns of polymorphism in *C. briggsae*. These authors pointed out that these trends are consistent with a more important role for BGS in self-fertilizing taxa and/or taxa with smaller effective population sizes (N_e) including humans (but see Enard et al. 2014) and *Arabidopsis thaliana*, compared to outcrossing species with large population sizes such as *Drosophila*, where positive selection seems to be the more predominant force. This relationship had also been pointed out earlier by Innan and Stephan (2003) who contrasted the relative roles of each type of selection in tomatoes (small effective population size) and *Drosophila* (large effective population size). Of course, the two forces are not mutually exclusive, and both forms of selection, positive and negative, likely act together to influence genomic evolution (Bustamante et al. 2005; Corbett-Detig et al. 2015).

This evidence that natural selection is an important force affecting evolution does not mean to say processes espoused by neutral theory – mutation, genetic drift and purifying selection – are not important themselves. And neutral theory itself has arguably provided a useful null model which makes clear, quantifiable predictions, and against which the presence of selection or other evolutionary forces can be tested. Rather, it has been suggested, that if natural selection is pervasive, then neutral theory should be replaced with a framework incorporating "*a much larger role for natural selection*" (Hahn 2008). Hahn (2008) suggested that the exact form that natural selection should take in this new 'selection theory' should depend on the biology of the species in question, based on evidence that the predominant form of selection varies from taxon to taxon. Cutter and Payseur (Cutter and Payseur 2013) suggest including BGS as part of null models in population genetics, and "*layering models of recurrent selective sweeps on top to determine whether and how much the incorporation of positive selection improves the*[ir] *fit*".

As well as the indirect effects of selection (eg. linkage to selected sites, as described above), sites that are often considered to be neutrally evolving, such as synonymous sites in coding sequences, introns, and intergenic sequences, may be directly affected by selection. For example, there is evidence in a variety of organisms for selection on codon usage (e.g. *Escherichia coli* (Sharp and Li 1986); *Mycobacterium tuberculosis* (Andersson and Sharp 1996), *Saccharomyces cerevisiae* (Sharp et al. 1986); various plants and algae (Morton 1998; Stenøien 2005); and *Drosophila* (Shields et al. 1988; Vicario et al. 2007)) whereby some codons out of the set that code for the same amino acid are preferred over the others. In *Drosophila*, where most preferred codons have a G or C nucleotide at the third position (Vicario et al. 2007), weak selection for preferred codons results in an increase in the GC content of synonymous sites (Shields et al. 1988; Moriyama and Hartl 1993). Other selective forces may also affect synonymous sites, for example to maintain splicing enhancers (Parmley et al. 2006), or mRNA secondary structure (Clarke 1970; e.g. Parsch et al. 1997; Carlini et al. 2001). These forces may be strong enough to mean that synonymous sites are non-neutrally evolving even in mammals, which have reasonably low effective population sizes and had historically been thought not to be affected by selection at such sites, which was assumed to be weak (Chamary et al. 2006).

Non-coding sequences, such as introns and intergenic regions, may also be directly selected (see Waterston et al. 2002; Lindblad-Toh et al. 2011). In *Drosophila*, indirect

evidence for this comes from the fact that non-coding DNA shows lower divergence but a higher skew towards rare alleles than synonymous sites in coding regions do, which suggests that non-coding regions are under greater selective constraint than synonymous sites (Andolfatto 2005). The same paper showed evidence for adaptive evolution in untranslated regions of mRNAs (UTRs) and, to a lesser extent, other non-coding regions including introns and intergenic regions (Andolfatto 2005). There may be at least three times the amount of functional, non-coding DNA as coding DNA in *Drosophila*, based on the number of sites inferred to be under selective constraint (Halligan and Keightley 2006). It seems likely that these functional non-coding sites play a role in gene regulation. One example of such an adaptive non-coding change is the insertion of transposable element fragments upstream of the *Cyp61g1* gene in *Drosophila melanogaster*, which increase its expression and so confer resistance to insecticides, including DDT. These resistance alleles are not present in pre-1940s samples but have subsequently spread to frequencies up to 100% in natural populations, coincident with the introduction of DDT (Daborn et al. 2002; Schmidt et al. 2010).

Non-selective but directional forces may also affect the composition of genomes, and may in turn be mistaken for signatures of demography or selection. One such force is mutation. For example, a positive correlation between the extent of codon usage bias (CUB) and expression level (per gene) is interpreted as evidence for selection for preferred codons, because more highly expressed genes should experience a greater selection pressure for translational efficiency (reviewed in Hershberg and Petrov 2008). If mutation drives the non-random use of codons then no such relationship is predicted. However, if mutational bias is associated with transcription, as it is in *Escherichia coli* and *Salmonella enterica* (Francino and Ochman 2001), then it is possible that a relationship between gene expression and base content could be driven by mutational bias, rather than selection. This possibility can be excluded by comparing intronic base composition, which would be affected by only the mutational bias, with synonymous base composition for the same gene, which would be affected by both mutational bias and selection (Duret 2002).

As a further example, it is currently unclear whether a reduction in CUB in the *D. melanogaster* species subgroup is attributable to a change in mutational bias, or a reduction in the scaled strength of selection ($4N_e s$, where N_e is the effective population size and s is the selection coefficient) favouring preferred codons. As mentioned above, selection for preferred codons in most *Drosophila* species results in an increase in GC

content at synonymous sites. However, the synonymous GC content in *D. melanogaster* has decreased since its common ancestor with *D. simulans*, as has the level of CUB (Akashi 1995; Akashi 1996; Kliman 1999; McVean and Vieira 2001; Akashi et al. 2006). Amongst the explanations for this pattern, the two best are 1) that mutation has shifted towards a greater level of AT bias or 2) that there has been a reduction in the scaled strength of selection on codon usage. Because the latter parameter is compound, either a reduction in the effective population size (N_e), perhaps due to a bottleneck (Akashi 1996), or a reduction in the selection coefficient (s), perhaps due to changing ecological conditions (Clemente and Vogl 2012a; Clemente and Vogl 2012b), may be responsible. Both a change in mutational bias towards AT, and a reduction in the scaled strength of selection for GC, are expected to perturb the mutation-selection base composition equilibrium towards a reduction in GC content at synonymous sites. Under neutrality, the fixation rate is equal to the mutation rate (Kimura 1983), so a shift towards a greater number of AT mutations leads to a greater number of AT fixations, resulting in a shift towards a lower GC content over time. If selection for GC becomes less effective (because of a reduction in $4N_e s$), the force opposing the AT mutational bias becomes weaker, and so also results in a lower GC content. Consequently, it is hard to tell these two processes apart using divergence data alone.

Another directional, but non-selective, force is GC-biased gene conversion (gBGC). gBGC is a recombination-associated process, which increases gametic GC content due to the way that the repair machinery associated with meiosis operates (Duret and Galtier 2009). gBGC, which can affect both substitution rates and the site frequency spectrum, may be mistaken for selection (Eyre-Walker 1999; Duret and Galtier 2009), and may act at any site at which recombination occurs. In *Drosophila*, (or any other organism) where preferred codons are GC-ending, gBGC may confound the detection of selection for preferred codons. Merely controlling for recombination as a co-variate may not distinguish the effects of gBGC and selection for preferred codons, because both are expected to be positively related to recombination (although, contrary to expectation, in *D. melanogaster* recombination rate and the extent of CUB are only weakly correlated on the autosomes, and are strongly negatively correlated on the X chromosome (Singh et al. 2005)). For selection on preferred codons, this is because in areas of high recombination, N_e, and so the strength scaled of selection, is higher (because of a reduction in the amount of interference between linked selected sites). In the case of gBGC, this is because it is a recombination-driven process. Further, although high

quality recombination information is available for *D. melanogaster* (Comeron et al. 2012) and some other model organisms, it is not for many other species, including other *Drosophila* species. The effects of these two processes may be separated by comparison between introns and synonymous sites in the same genes, because the former class of site is expected to be affected by gBGC but not selection for preferred codons, and the latter by both (Zeng and Charlesworth 2010a). However, as we have seen above, introns themselves may be subject to other constraints (or positive selective forces), complicating matters. In *Drosophila*, the 8-30bp region of short (less than 66bp long) introns seems to be most neutrally evolving genomic region, based on polymorphism and divergence data (Halligan and Keightley 2006; Parsch et al. 2010; Clemente and Vogl 2012b), which means that these sites may be used as a reference against which the action of CUB and other selective forces can be tested.

Finally, the patterns of polymorphism expected under natural selection may also be confounded by demographic changes (Tajima 1989a; Tajima 1989b). In the case of Tajima's *D* statistic (Tajima 1989a), which is based on the comparison of the number of segregating sites and the number of pairwise differences in a sample of sequences, both a departure from neutrality (Tajima 1989a) and demographic events, for example a bottleneck (Tajima 1989b) may result in a significant test statistic.

"*An unconstrained reference sequence facilitates the detection of selection*"

(Clemente and Vogl 2012b)

Disentangling these many selective and non-selective forces in order to understand how each contributes to evolution is an important undertaking. One approach, which is touched on above, is to compare different classes of site thought to be subject to different evolutionary pressures, in order to infer the processes acting at each. This is the approach taken by the McDonald-Kreitman test, for example (McDonald and Kreitman 1991). In the original formulation of this test, synonymous and non-synonymous divergence and polymorphism are compared to detect the action of selection at the non-synonymous sites. However, as we have also seen above, neutrally evolving sequences may not be as common as previously assumed, and this has implications for the McDonald-Kreitman framework – for instance, in the presence of weakly deleterious mutations estimates of α are downwardly biased (Eyre-Walker and Keightley 2009). Nevertheless, this comparative method offers a powerful framework

and is widely used (Welch 2006). In a similar vein, it may also be illuminating to compare the same class of sites (e.g. synonymous, non-synonymous, intronic, intergenic, etc.) but which differ in other factors which are expected to affect evolution, for instance mode of inheritance (e.g. autosomal or sex-linked), recombination rate, and selective constraint. For example, as we saw above, the relationship between recombination rate and genetic diversity found in all organisms assayed to date is the hallmark of pervasive natural selection. In *Drosophila*, the ratio of N_e between the X chromosome and the autosomes might be assumed to be 3/4, which suggests that X-linked synonymous diversity should be 3/4 of the value of autosomal diversity. However, the observed diversity ratio is close to one, which may be explained by the fact that the effective recombination rate is higher on the X (by 4/3) because males lack crossing over. This can be seen from the fact that the X chromosome spends 2/3 of its time in females (which are XX), and 1/3 of its time in males (which are XY), whereas the autosomes spend an equal amount (1/2) of their time in each sex. As there is no crossing over in males, the effective recombination rate for the autosomes is 1/2*r*, whereas for the X it is 2/3*r*, where *r* is a given recombination rate in females. This results in less polymorphism being removed by pervasive linked selection on the X compared to the autosomes (Charlesworth 2012a).

Genomic or detailed genetic data are now increasingly available for both model and non-model organisms, and this is allowing researchers to further investigate the role that different forces play in shaping the evolution of species' genomes (Stapley et al. 2010). In this thesis I use large genetic and genomic datasets to attempt to quantify the action of natural selection and other forces on patterns of genetic polymorphism, differentiation and divergence.

Outline of data chapters

In Chapter 2, I ask to what extent purifying selection modulates differentiation between populations of *Drosophila melanogaster*. This question is pertinent because F_{ST} genome scans are often employed to detect candidate loci linked to the targets of positive selection, but such tests are not designed to take into account pervasive selection (Bierne et al. 2013). Both selective sweeps and BGS reduce variation at sites linked to selected loci, which will have the effect of reducing within-population diversity relative to between-population diversity, and so inflating relative measures of differentiation

such as F_{ST} (Charlesworth et al. 1997; Charlesworth 1998; Cruickshank and Hahn 2014; Zeng and Corcoran 2015). Additionally, because the effect of selection on linked sites is modulated by recombination, it is important to take recombinational context into account. It is easy to imagine a scenario in which the signature of a selective sweep event in a region of high recombination is quickly broken down, whilst the signature left by an equally advantageous sweep in a region of low recombination persists for longer and therefore is more likely to be detected (Roesti et al. 2012). Alternatively, one could imagine that the action of purifying selection on many sites in a region of low recombination might result in similar patterns to a single selective sweep in a different recombinational context (Charlesworth et al. 1997). Consequently, it is not clear to what extent BGS might confound the identification of positively-selected sites by F_{ST} outlier methods. I answer this question using high quality whole-genome data for a pair of populations of *D. melanogaster* – one French and one Rwandan – published by the Drosophila Population Genomics Project (Langley et al. 2012; Pool et al. 2012). Novel findings from this analysis include the fact that that purifying selection affects population differentiation, as measured by F_{ST}, at 0-fold degenerate and intronic sites, via its impact on minor allele frequency. Further, I uncover a positive relationship between intron length and selective constraint in *D. melanogaster* as well as some important statistical properties of F_{ST}, which bear on its ability to detect signals of selection.

In Chapter 3, I investigate patterns of polymorphism and divergence at a locus underlying a colour phenotype in the Australian finch *Erythrura gouldiae*, with a view to determining whether natural selection has acted at this locus. Previous work had mapped a Mendelian locus controlling cheek patch colour in these birds to a regulatory region on the Z chromosome, named the *Red* locus (Kim et al, in prep). The colour phenotype which the *Red* locus underlies – red versus black cheek patches – is implicated in several behavioural and genetic traits, including behavioural dominance (Pryke and Griffith 2006; Pryke 2007), mate choice (Pryke and Griffith 2007; Pryke and Griffith 2009a) and post-zygotic isolating mechanisms (Pryke and Griffith 2009b). Both red and black-cheeked birds co-occur in natural populations, at frequencies which seem to be stable over time (Pryke and Griffith 2009b). This suggests that the *Red* locus might be subject to some form of natural selection. My analyses of Z-linked Gouldian Finch data allows me to reject the null hypothesis of neutrality at the *Red* locus, where the observed patterns of polymorphism and divergence are most compatible with the

action of long-term balancing selection. To reach this conclusion, I account for important population genetic parameters including the population mutation rate, the population recombination rate, and demographic changes, as well as a non-random sampling scheme which resulted in non-natural allele frequencies in our sample of the *Red* locus. This sampling scheme had implications for the statistical power of some tests, and also rendered many standard population genetics approaches inappropriate, which meant that tailored methods were required. The methods used include coalescent simulations and an implementation of HKA test, which was extended to account for non-random sampling, recombination, and unequal locus lengths.

In Chapter 4, I investigate the extent to which selection for preferred codons affects evolution at 4-fold degenerate sites in *Drosophila simulans* and *D. melanogaster*. Quantifying selection for preferred codons is important because it affects a large proportion of the genome, and acts on sites which are often used as a 'neutral' reference against which selection at other sites is inferred, for example using the McDonald-Kreitman test (McDonald and Kreitman 1991; above). It follows that quantifying the effect of selection on these sites is important for the accurate inference of natural selection elsewhere in the genome. It may also have implications for the amount of genetic load that populations experience. Previous work on this topic in *D. melanogaster* was complicated by an excessive AT fixation bias in the *D. melanogaster* lineage (Akashi 1995; Akashi 1996; Poh et al. 2012), and by the use of methods which do not take this non-equilibrium situation into account. As a consequence of these difficulties, there is currently some uncertainty about i) the current strength of selection for preferred codons in *D. melanogaster*; ii) the timing of changes in the strength of selection compared to the common ancestor of *D. melanogaster* and *D. simulans*; iii) the nature of the forces underlying these changes. Less is known about the extent of CUB in *D. simulans*, and a comparison between the two species is useful in helping us understand the differences between them, as well as the forces acting in each. To this end, I obtain new whole polymorphism data from an Africa population of *D. simulans* and compare it to the African population of *D. melanogaster* described in Chapter 2, using *D. yakuba* as an outgroup. I find evidence of longer term and more recent (within $4N_e$ generations) selection for preferred codons in both species, with the evidence for recent selection being much stronger in *D. simulans*.

Chapter 2. The effects of purifying selection on patterns of genetic differentiation between *Drosophila melanogaster* populations

Contributing authors: Benjamin C. Jackson, José L. Campos and Kai Zeng

This paper is published in: Heredity **114**: 163-174. doi:10.1038/hdy.2014.80

KZ and JLC agree to incorporation of this work in this thesis. I have retained the right to publish this material as part of this thesis.

Abstract

Using data provided by the *Drosophila* Population Genomics Project, we investigate factors that affect genetic differentiation between Rwandan and French populations of *D. melanogaster*. By examining within-population polymorphisms, we show that sites in long introns (especially those > 2000bp) have significantly lower π (nucleotide diversity) and more low-frequency variants (as measured by Tajima's *D*, minor allele frequencies, and prevalence of variants that are private to one of the two populations) than short introns, suggesting an positive relationship between intron length and selective constraint. A similar analysis of protein-coding polymorphisms shows that 0-fold (degenerate) sites in more conserved genes are under stronger purifying selection than those in less conserved genes. There is limited evidence that selection on codon bias has an effect on differentiation (as measured by F_{ST}) at 4-fold (degenerate) sites, and 4-fold sites and sites in 8-30bp of short introns ≤ 65bp have comparable F_{ST} values. Consistent with the expected effect of purifying selection, sites in long introns and 0-fold sites in conserved genes are less differentiated than those in short introns and less conserved genes, respectively. Genes in non-crossover regions (e.g., the 4th chromosome) have very high F_{ST} values at both 0-fold and 4-fold degenerate sites, which is probably due to the large reduction in within-population diversity caused by tight linkage between many selected sites. Our analyses also reveal subtle statistical properties of F_{ST}, which arise when information from multiple SNPs is combined and can lead to the masking of important signals of selection.

Introduction

Natural populations are often divided into subpopulations. Studying the extent to which different subpopulations are genetically differentiated has been of paramount importance in evolutionary genetics, as it provides a way to examine how different evolutionary forces such as genetic drift, natural selection, and migration drive changes in the genome (reviewed in Chap. 7 of Charlesworth and Charlesworth 2010).

Specifically, insights into fundamental processes such as historical demographic changes, (local) adaptation, and speciation can be obtained by comparing patterns of genetic differentiation across different genomic regions (Wu 2001; Weir and Hill 2002; Charlesworth et al. 2003; Hey and Machado 2003; Beaumont 2005; Holsinger and Weir 2009). For instance, by scanning for loci that show unusually high levels of differentiation relative to the rest of the genome, we can detect loci that are under diversifying selection, whereby different alleles are favoured in different subpopulations (Beaumont and Nichols 1996; Beaumont and Balding 2004; Foll and Gaggiotti 2008; Excoffier et al. 2009). As another example, in a study comparing African and non-African humans, it was found that the X chromosome was substantially more diverged than the autosomes, over and above the null expectation based on the fact that there are four copies of each autosome for every three copies of the X chromosome, which in turn suggests that dispersal in humans may be sex-biased or that the X chromosome may have experienced repeated selection after the divergence of African and non-African populations (Keinan et al. 2009).

Genetic differentiation between subpopulations is often measured by Wright's F_{ST} (Wright 1951), which is abbreviated as F in this study. F can be defined as the proportion of genetic variation explained by differences in allele frequencies between subpopulations (Charlesworth 1998; Holsinger and Weir 2009; Bhatia et al. 2013). F ranges between 0 and 1, which indicate no differentiation and fixed differences between subpopulations, respectively. Various genetic data, e.g. single nucleotide polymorphisms (SNPs) and microsatellites, can be used to estimate F, but using statistical procedures that take into account biological properties of the data under consideration (e.g., high versus low mutation rate) is vital for acquiring accurate estimates (Weir and Cockerham 1984; Excoffier et al. 1992; Slatkin 1995; Nagylaki 1998; Holsinger and Weir 2009).

Drosophila melanogaster, a classic model organism for population genetics, offers an invaluable system for studying population differentiation. Despite having a worldwide distribution in the present day, it is believed that the species originated in sub-Saharan Africa (David and Capy 1988; Stephan and Li 2007). The colonisation of Europe has been suggested to have taken place about fifteen thousand years ago (David and Capy 1988; Stephan and Li 2007; Duchen et al. 2013). The Americas and Australia were colonised much more recently, possibly in the past few hundred years (David and Capy 1988; Stephan and Li 2007; Duchen et al. 2013). By studying patterns of genetic differentiation, investigators have obtained evidence that American populations of *D. melanogaster* may be formed by admixture between African and European flies (Caracristi and Schlötterer 2003). Multiple attempts have also been made to identify loci with unusually high F_{ST} which may have contributed to local adaptation to different habitats (Turner et al. 2008; Yukilevich et al. 2010; Kolaczkowski et al. 2011; Fabian et al. 2012; Langley et al. 2012; Pool et al. 2012; Campo et al. 2013).

These previous studies of *D. melanogaster* have mainly focused on determining the evolutionary relationship between subpopulations, quantifying the overall level of differentiation, and detecting genomic regions of interest using outlier scans. However, the role of purifying selection in shaping large-scale patterns of differentiation has not been well characterised, although it has been widely accepted that the majority of new mutations that affect fitness will have detrimental effects (Pál et al. 2006; Eyre-Walker and Keightley 2007). Supporting this view, it has been estimated that only between 1% and 2% of new nonsynonymous mutations in *D. melanogaster* are (weakly) positively selected, and about 6% are nearly neutral (i.e., $|N_e s| \leq 1$, where N_e is the effective population size and s the selection coefficient), and the remaining are deleterious ($|N_e s| > 1$) (Eyre-Walker and Keightley 2007; Eyre-Walker and Keightley 2009; Schneider et al. 2011). Thus, we are interested in testing the following predictions based on population genetic theory of subdivided populations (reviewed in Chap. 7 of Charlesworth and Charlesworth 2010): (*i*) purifying selection reduces differentiation between populations at functionally important regions; (*ii*) the level of reduction is positively correlated with the level of selective constraint. Answering these questions will help us better understand the sources of variation in genetic differentiation across the genome, which is important for example, in interpreting results obtained from genome scans (Beaumont and Nichols 1996; Beaumont and Balding 2004; Foll and Gaggiotti 2008; Excoffier et al. 2009).

We will address the questions raised above by making use of the high-quality whole-genome resequencing data published by the *Drosophila* Population Genomics Project (DPGP) for one French population and one Rwandan population (Langley et al. 2012; Pool et al. 2012). In addition to protein-coding regions, we investigate introns, as previous studies have shown strong evidence that these genomic regions are under substantial selective constraints, probably as a result of the presence of *cis*-regulatory elements and noncoding RNA genes (Bergman and Kreitman 2001; Parsch 2003; Andolfatto 2005; Haddrill et al. 2005; Halligan and Keightley 2006; Casillas et al. 2007; Roy et al. 2010).

Our study proceeds as follows. First, we present an overview of patterns of genetic variation both within and between populations using data from genomic regions where crossing over occurs (crossover regions). We are interested in understanding whether patterns of differentiation at 4-fold degenerate (hereafter 4-fold) sites are affected by selection on codon usage, and whether 4-fold sites and putatively neutral sites in 8-30 bp regions of introns ≤ 65 bp (Halligan and Keightley 2006; Parsch et al. 2010) are comparable with respect to levels of differentiation. These are intended to identify putatively neutral sites which can be used as a reference in the study of the effects of purifying selection on genetic differentiation. We then examine the relationship between K_A (nonsynonymous divergence) and diversity/differentiation patterns at 0-fold degenerate (hereafter 0-fold) sites in protein-coding regions, as well as the relationship between intron length and diversity/differentiation patterns in intronic regions. Finally, we compare non-crossover regions (e.g., the fourth chromosome) and crossover regions regarding differentiation patterns, study the relative contribution of selection and genetic linkage, and examine the correlation between local recombination rates and F at putatively neutral sites.

Materials and Methods

Data acquisition

To obtain polymorphism and divergence data we downloaded FASTQ files from the *Drosophila* Population Genomics Project (http://www.dpgp.org/dpgp2/candidate/) for 17 Rwandan *D. melanogaster* samples (RG18N, RG19, RG2, RG22, RG24, RG25, RG28, RG3, RG32N, RG33, RG34, RG36, RG38N, RG4N, RG5, RG7 and RG9), which have been estimated to have the lowest estimated levels of admixture with

European populations (less than 3%, see Figure 3B of Pool et al. 2012). We also selected seven samples from the French population (FR14, FR151, FR180, FR207, FR217, FR310 and FR361). We will refer to these two samples as RG and FR, respectively. These samples are from inbred isofemale lines, and are haploid due to the loss of heterozygosity during the inbreeding process. We further masked any regions of the African samples with evidence of admixture from European populations, using the admixture coordinates reported by Pool et al. (2012). Sites with a quality score below 31 (equivalent to a PHRED score of 48, and approximately equivalent to one error per 100kb; see Pool et al. 2012) were also masked.

From the FASTQ files we extracted protein-coding regions in crossover regions, which we abbreviate as CDS-C, using gene annotations from FlyBase release version 5.44 (www.flybase.org) and made FASTA files containing all samples (24 alleles). For each *D. melanogaster* gene with multiple transcripts, we chose one transcript randomly.

Introns belonging to our chosen transcript were extracted, and were further processed by masking any coding regions that form part of the other transcripts. Only introns occurring in crossover regions were retained (polymorphism data for introns in non-crossover regions were of significantly lower quality, and were therefore excluded).

Protein-coding regions in non-crossover regions of the *D. melanogaster* genome, abbreviated as CDS-NC, were obtained from Campos et al (2012). These data included five unlinked regions: N2 (genes located in heterochromatic regions near the centromere of the second chromosome), N3 (genes located in heterochromatic regions near the centromere of the third chromosome), N4 (the fourth chromosome), NXc (X-linked genes near the centromere), and NXt (X-linked genes near the telomere).

For all CDS-C and CDS-NC we selected *D. yakuba* as an outgroup to avoid any major influence of ancestral polymorphisms on the estimation of sequence divergence, which can potentially create spurious correlations between sequence divergence and recombination (e.g., Cutter and Choi 2010). One-to-one orthologous *D. yakuba* sequences were obtained from FlyBase (available at ftp://ftp.flybase.net/releases/FB2012_02/precomputed_files/genes/gene_orthologs_fb_2012_02.tsv.gz). We then performed amino-acid sequence alignments using MAFFT (Katoh et al. 2002). These amino-acid sequence alignments were translated back to nucleotides using custom scripts in PERL to produce in-frame coding sequence alignments that included the 24 *D. melanogaster* alleles and the *D. yakuba* reference.

For introns we used *D. simulans* as an outgroup since we considered *D. yakuba* too distant for producing reliable alignments, due to the increased prevalence of indels in non-coding regions. We obtained orthologous intronic *D. simulans* sequences from Hu et al. (2013), which was based on an updated *D. simulans* genome assembly and careful alignment procedures to preserve gene structures (http://genomics.princeton.edu/AndolfattoLab/w501_genome_files/alnMSY.tar.gz).

Recombination rate for the midpoint of all alignments was obtained using the *Drosophila melanogaster* Recombination Rate Calculator v2.3 (Fiston-Lavier et al. 2010) and the high-resolution genetic map published recently by Comeron et al. (2012).

Sequence Analysis

For CDS-C and CDS-NC, we calculated K_A and K_S (the numbers of nonsynonymous and synonymous substitutions per nonsynonymous and synonymous site, respectively) using the kaks() function from the seqinr package (Charif and Lobry 2007) in R (http://www.r-project.org/), which implements the method of Li (1993). For introns, we calculated divergence (*K*) to the *D. simulans* reference using the dist.dna() function in the ape package of R (Paradis et al. 2004), with the 'K80' method (Kimura 1980). For conducting analyses using polymorphism data in the two *D. melanogaster* samples, we split CDS-C and CDS-NC into 0-fold degenerate sites and 4-fold degenerate sites by analysing the alignments codon by codon. A codon column was retained if the following requirements were met: (*i*) data from all individuals were available; (*ii*) it had at most one SNP. These were to avoid uncertainty of the order of mutations in codons with multiple SNPs. We retained 7235 autosomal and 1150 X-linked CDS alignments for which we had both more than 10 bp of 0-fold sites and more than 10bp of 4-fold sites.

We split introns into short (≤ 65bp long) and long (> 65bp long) classes, and further trimmed short introns to retain positions 8-30 from the 5' end (the SI sites), in order to retain sites under the least amount of selective constraint (Halligan and Keightley 2006; Parsch et al. 2010). This left us with 7483 autosomal and 752 X-linked short introns, and 8851 autosomal and 1869 X-linked long introns. To keep the sample size the same as the CDS data, only intronic sites where data from all individuals were available were retained. This requirement appears to be conservative with respect to detecting SNPs (note also that regions within 3 bp of an indel were also masked by DPGP; Pool et al.

2012 and http://www.dpgp.org/dpgp2/DPGP2.html). For instance, π (nucleotide diversity) estimated using data from SI sites that fulfilled the above criterion was 0.0146 in the RG sample, whereas the estimate increased to 0.0164 when we retained, within the same genomic regions, all sites that had data from at least two individuals. However, as we will show in the Results, there is no detectable difference between the SI and 4-fold sites in terms of skewness of allele frequency spectrum (as measured by Tajima's *D*), average minor allele frequency, and F_{ST}. Our conservative data filtering procedure is unlikely to bias our analysis of population differentiation.

Because of the complete linkage between the CDS-NC genes located within a non-crossover (NC) region, each non-recombining region of the genome effectively represents a single locus. Therefore all genes within a single NC region were concatenated and analysed as a single gene. In total, we have three autosomal NC regions, which are N2, N3, and N4, and two X-linked NC regions, which are NXc and NXt. These loci were kept intact in the permutation tests used to compare values of summary statistics calculated using data from NC and crossover (C) regions.

Nucleotide diversity (π), Tajima's *D* (Tajima 1989a), and relative Tajima's *D*, (Schaeffer 2002) were calculated using nuc.div() and a modified version of the tajima.test(), both from the pegas package in R (Paradis 2010). Since conclusions drawn from Tajima's *D* and relative Tajima's *D* are identical, only the former is presented. Permutation tests were carried out to assess whether these statistics were different between different types of sites. For instance, to compare whether values of Tajima's *D* at 4-fold sites and SI sites were comparable, 10,000 pseudosamples were generated by randomly shuffling both the 4-fold and SI sites in the data (SNPs from a single locus were shuffled as a unit), such that in each pseudosample there were similar numbers of SNPs in the two "site classes" as in the real data.

To assess the effects of selection on codon bias on population differentiation at 4-fold sites, we calculated Fop (frequency of optimal codons) with CodonW (Peden 1999), using the built-in table of optimal codons for *D. melanogaster*.

Measuring population differentiation

Levels of differentiation between the two populations of *D. melanogaster* were measured by Wright's F_{ST}, which is abbreviated as F. We used the definition of Weir and Cockerham (1984), which can be expressed as

$$F = \frac{\pi_B - \pi_S}{\pi_B} \quad (1)$$

where π_B is the expected divergence between a pair of alleles sampled from two different populations, and π_S is the expected within-population diversity (see also Charlesworth 1998; Keinan et al. 2007). Previous investigations (Maruki et al. 2012; Jakobsson et al. 2013) of the dependence of F on minor allele frequency (MAF) were based on a different definition of F put forward by Nei (1973). We therefore derive the maximum value of Weir and Cockerham's definition of F as a function of MAF.

Consider a population divided into two subpopulations. We examine a biallelic locus. The frequency of one of the two alleles in the k-th subpopulation is referred to as p_k (k = 1 or 2). It can be shown that $\pi_B = p_1(1 - p_2) + p_2(1 - p_1)$ and $\pi_S = p_1(1 - p_1) + p_2(1 - p_2)$ (Charlesworth 1998). Substituting these into Eq. (1), F can be rewritten as

$$F = \frac{2\delta^2}{2\sigma - \sigma^2 + \delta^2} \quad (2)$$

where $\delta = |p_1 - p_2|$ and $\sigma = p_1 + p_2$. Without loss of generality, we assume that $1 \leq \sigma \leq 2$. Three properties are of use: (*i*) MAF = $1 - \sigma/2$; (*ii*) $0 \leq 2\sigma - \sigma^2 \leq 1$ for $1 \leq \sigma \leq 2$; (*iii*) $0 \leq \delta \leq (2 - \sigma)^2$. Rearranging Eq. (2), we deduce that

$$F = 2\left[1 - \frac{2\sigma - \sigma^2}{2\sigma - \sigma^2 + \delta^2}\right] \leq 2\left[1 - \frac{\sigma(2 - \sigma)}{\sigma(2 - \sigma) + (2 - \sigma)^2}\right] = 2 - \sigma \quad (3)$$

Since MAF = $1 - \sigma/2$, the above inequality is equivalent to max(F) ≤ 2 MAF. In Supplementary Figure S1, we display the differences between the upper bounds of F derived here and that obtained in previous studies using Nei's F (Maruki et al. 2012; Jakobsson et al. 2013). It can be seen that F can only assume a very restrictive range of values when MAF is small.

The estimator of F proposed by Hudson et al. (1992; see also Keinan et al. 2007; Bhatia et al. 2013) was employed:

$$\hat{F} = \frac{\hat{\pi}_B - \hat{\pi}_S}{\hat{\pi}_B} \tag{4}$$

where $\hat{\pi}_B$ and $\hat{\pi}_S$ are estimates of π_B and π_S obtained from data. Eq. (4) can be calculated using information from a single SNP. To combine information from multiple SNPs, the following two methods were used (Weir and Cockerham 1984; Bhatia et al. 2013):

$$F^U = \frac{1}{S}\sum_{i=1}^{S} \hat{F}^{(i)} \tag{5}$$

and

$$F^W = \frac{\sum_{i=1}^{S}\left(\hat{\pi}_B^{(i)} - \hat{\pi}_S^{(i)}\right)}{\sum_{i=1}^{S} \hat{\pi}_B^{(i)}} \tag{6}$$

where S is the number of SNPs, and $\hat{F}^{(i)}$, $\hat{\pi}_B^{(i)}$ and $\hat{\pi}_S^{(i)}$ are values of the terms defined in Eq. (4) obtained using data from the i-th SNP.

It should be noted that F^U gives equal weight to all SNPs, whereas F^W gives more weight to SNPs with higher expected levels of polymorphism. In other words, F^U is expected to be more sensitive to the presence of SNPs with low MAFs, but F^W is dominated by SNPs that are on average more polymorphic. To see this more explicitly, assume that we are combining information from two SNPs (i.e., $S = 2$). We add a subscript j to the symbols defined above to signify the locus under consideration, so that we have p_{jk}, σ_j, and $\delta_j = |p_{j1} - p_{j2}|$. We further assume that $2 > \sigma_1 \geq \sigma_2 \geq 1$. Note that σ_j are regarded as parameters (e.g., a SNP under strong selective constraints is expected to have a larger σ [i.e., a smaller MAF] than a neutral SNP). Some straightforward algebra leads to the following results: (*i*) $\max[F(\sigma_1)] \leq \max[F^U] \leq \max[F(\sigma_2)]$; (*ii*) $\max[F(\sigma_1)] \leq \max[F^W] \leq \max[F(\sigma_2)]$, where $\max[F(\sigma_j)] = 2 - \sigma_j$ (these results hold when $S > 2$; proof not shown). To see the differential sensitivities to SNPs with small MAFs, we

define $\Delta_1(U) = \max[F^U] - \max[F(\sigma_1)]$, $\Delta_2(U) = \max[F(\sigma_2)] - \max[F^U]$, $\Delta_1(W) = \max[F^W] - \max[F(\sigma_1)]$, and $\Delta_2(W) = \max[F(\sigma_2)] - \max[F^W]$. Using Eq. (3), we show that $\Delta_1(U)/\Delta_2(U) = 1$, but $\Delta_1(W)/\Delta_2(W) = (2 - \sigma_2)/(2 - \sigma_1) \geq 1$. Thus, the behaviour of F^W is more akin to that of the more polymorphic SNP (i.e., $\max[F^W]$ is closer to $\max[F(\sigma_2)]$ than to $\max[F(\sigma_1)]$). As we will see later, this property of F^W can lead to the masking of important signatures of evolution when SNPs with different properties are combined.

Results

Genome-wide polymorphism patterns in crossover regions

Table 1 presents summaries of polymorphism patterns for autosomal (A) and X-linked (X) loci situated in genomic regions where crossing-over occurs (crossover regions). For ease of presentation, we will refer to nucleotide diversity, π, calculated using 0-fold sites, 4-fold sites, and SI sites (positions 8-30 from the 5' end of short introns ≤ 65bp) as π_0, π_4, and π_{SI}, respectively; a similar notational convention will be used for other statistics. For both A and X, and in both the Rwandan (RG) and French (FR) samples, π_0, Tajima's D_0 (Tajima 1989a), and MAF_0 (minor allele frequency) are significantly smaller than the corresponding estimates obtained from 4-fold and SI sites ($P_{\text{permutation}} < 0.001$ in all cases), consistent with the well-known fact that most nonsynonymous mutations are deleterious (Pál et al. 2006; Eyre-Walker and Keightley 2007), and are therefore kept at low frequencies in the population by purifying selection (Kimura 1983). Previous studies have suggested that SI sites may be neutrally evolving (Halligan and Keightley 2006; Parsch et al. 2010). In our data set, π_{SI} seems to be somewhat smaller than π_4, which may be due to the stringent data filtering procedure we employed (see Materials and Methods), or the higher GC content at 4-fold sites compared to intronic sites, which in turn is expected to result in an increased mutation rate in 4-fold sites (Singh et al. 2005; Keightley et al. 2009). There is, however, no statistically discernible difference with respect to either MAF or Tajima's D between 4-fold and SI sites (Table 1; $P_{\text{permutation}} > 0.1$ for both A and X).

The FR sample has a lower level of diversity than RG for all three types of sites (Table 1), reflecting a loss of genetic variation induced by population bottlenecks which are believed to have occurred as the species migrated out of Africa (Haddrill, Thornton, et

al. 2005; Li and Stephan 2006; Thornton and Andolfatto 2006; Hutter et al. 2007; Duchen et al. 2013). The difference in π_0 between the two populations is somewhat smaller than those observed for π_4 and π_{SI} [e.g., on A, π_0(FR)/π_0(RG) = 0.83 *versus* π_4(FR)/π_4(RG) = 0.77]. This is probably because more 0-fold sites are under strong selective constraint, so that variants at these sites behave almost deterministically, and are therefore less sensitive to demographic changes (Zeng 2013).

To inspect overall patterns of genetic differentiation between the RG and FR populations, we calculated F_{ST} [abbreviated here as F; see Eq. (1) in Materials and Methods], as defined by Weir and Cockerham (1984), using the estimator of Hudson et al. (1992). Two approaches were employed to combine information over multiple SNPs: un-weighted mean F [Eq. (5)] and weighted mean F [Eq. (6)], which will be referred to as F^U and F^W, respectively. Since most nonsynonymous mutations are likely to be deleterious, it is expected that levels of population differentiation at these selectively constrained sites should be lower than those at less constrained sites (e.g., 4-fold sites) (Barreiro et al. 2008; Maruki et al. 2012). Surprisingly, values of F_0^W, estimated using either the autosomal or X-linked data, are not statistically different from those of either F_4^W or F_{SI}^W (Table 1; $P_{\text{permutation}} > 0.1$ in all cases). There is also no detectable difference between F_4^W and F_{SI}^W ($P_{\text{permutation}} > 0.1$ for both A and X). In contrast, F_0^U was found to be significantly smaller than both F_4^U and F_{SI}^U ($P_{\text{permutation}} < 0.001$ for both A and X), while the differences between F_4^U and F_{SI}^U remain non-significant ($P_{\text{permutation}} > 0.1$ for both A and X). The patterns obtained from F^U are therefore more compatible with the *a priori* expectation that 0-fold sites are on average more constrained than 4-fold and SI sites. We will investigate causes for the lack of difference between F_0^W and either F_4^W or F_{SI}^W in a later section.

Several differences between A and X are of note (Table 1). Firstly, consistent with previous reports (Caracristi and Schlötterer 2003; Hutter et al. 2007; Charlesworth 2012a; Pool et al. 2012; Campos et al. 2013), the X:A ratio in diversity at putatively neutral sites (i.e., 4-fold and SI sites) is about 1 in the RG population [π_4(X)/π_4(A) = 1.08 and π_{SI}(X)/π_{SI}(A) = 1.10], higher than the *null* expectation of 3/4. Secondly, the reduction in diversity in FR is more pronounced for X than A for all three types of sites [e.g., π_4(FR)/π_4(RG) = 0.41 and 0.77 for X and A, respectively], as reported in previous

investigations (Caracristi and Schlötterer 2003; Hutter et al. 2007). Finally, the extent of population differentiation at both 4-fold and SI sites, as measured by either F^U or F^W, is significantly higher on the X than on A ($P_{\text{permutation}} < 0.001$ for all comparisons). This is probably largely driven by the greater reduction in diversity on the X in non-African populations, as values of D_{xy}, the mean number of nucleotide substitutions between sequences taken from different subpopulations (Nei and Miller 1990), are comparable between A and X in this study: $D_{xy,4}$ = 1.65% and 1.64%, and $D_{xy,SI}$ = 1.51% and 1.58%. A systematic examination of possible causes of the apparent differences between A and X is beyond the scope of this study; the interested reader can refer to previous studies of this topic (Charlesworth 2001; Pool and Nielsen 2007; Singh et al. 2007; Pool and Nielsen 2008; Yukilevich et al. 2010; Charlesworth 2012a; Campos et al. 2013). In what follows, results obtained from A and X will be presented separately.

Limited evidence for selection on codon usage bias affecting patterns of population differentiation at 4-fold degenerate sites

To investigate whether selection on codon usage bias (CUB) affects differentiation patterns at 4-fold sites, we first examined the relationship between F_4^U and *Fop* (frequency of optimal codons), as the latter is well known to be correlated with the intensity of selection on CUB (reviewed in Hershberg and Petrov 2008; Zeng and Charlesworth 2009). Considering the large variance of the *F* estimators and the dearth of SNPs in individual genes, we grouped the genes into equal-sized bins with similar numbers of SNPs at 4-fold sites. As shown in Supplementary Figure S2A, *Fop* and F_4^U are not correlated on A (Kendall's τ = -0.01, $P > 0.1$). On the X, some evidence for a weak negative correlation was obtained (Figure S2B), but it is not statistically significant (Kendall's τ = -0.6, $P = 0.13$). When F_4^W was considered, no correlation was found on either A or X (Supplementary Figure S2E and S2F). To investigate this further, for the genes within each bin on the X, we tested whether F_4^U differed from F_{SI}^U statistically. Amongst the six bins, no evidence of a significant difference was found for the first four bins, whereas the differences were marginally significant for the last two bins with highest *Fop* ($P_{\text{permutation}}$ = 0.04 and 0.05, respectively). Similarly, we

did not detect any correlation between K_S and either F_4^U or F_4^W (Supplementary Figure S2).

Overall, there is limited evidence that selection on CUB is strong enough to substantially alter patterns of genetic differentiation at 4-fold sites. Considering that 4-fold and SI sites in crossover regions are comparable with respect to both MAF and *F*, in what follows, we will use population differentiation patterns obtained from the two types of site as neutral standards, and will refer to them as putatively neutral sites.

Evolutionarily conserved genes are under stronger purifying selection and have reduced F at 0-fold degenerate sites

Genes in crossover regions were divided into equal-sized bins (with similar numbers of SNPs) based on their K_A values between *D. melanogaster* and *D. yakuba*. We inspected polymorphism patterns in the RG sample as a function of K_A; a qualitatively identical set of results were obtained using the FR sample (Supplementary Figure S3). On both A and X, K_A was found to be significantly positively correlated with both π_0 (Figures 1A and 1B; A: Kendall's $\tau = 0.989$ and $P < 0.001$; X: Kendall's $\tau = 1$ and $P = 0.009$) and Tajima's D_0 (Figures 1C and 1D; A: Kendall's $\tau = 0.884$, $P < 0.001$; X: Kendall's $\tau = 0.867$ and $P = 0.024$). No statistically significant relationship was found when comparing K_A with Tajima's D_4 (Figures 1C and 1D; Kendall's $\tau = -0.2$ and -0.333, $P > 0.1$, for X and A), although there is a negative correlation between K_A and π_4 on A (Figure 1A; Kendall's $\tau = -0.6$, $P < 0.001$) (see also Andolfatto, 2007; Haddrill *et al.*, 2011). In particular, on both A and X, π_0 and Tajima's D_0 approach π_4 and Tajima's D_4, respectively, as K_A increases. In contrast, values of π_4 and Tajima's D_4, regardless of the K_A bin from which they were obtained, remain similar to the values of π_{SI} and Tajima's D_{SI}. These results suggest that 0-fold sites are under stronger constraints than 4-fold and SI sites, and that 0-fold sites in genes with smaller K_A are, on average, under stronger purifying selection. We obtained the same results when we used the *D. simulans* genome as an outgroup (Supplementary Figure S4).

Figures 2A and 2B show that evolutionarily conserved genes have significantly smaller F_0^U (A: Kendall's $\tau = 0.663$, $P < 0.001$; X: Kendall's $\tau = 0.867$, $P = 0.02$). Again, we obtained the same result when using *D. simulans* as the outgroup (Supplementary

Figure S5). The pattern remains statistically significant for autosomes when F_0^W was considered (Supplementary Figure S6). The reduction in F_0 for genes with smaller K_A is associated with a strong reduction in MAF_0 (Figures 2C and 2D) and an increase in the proportion of 0-fold SNPs that are private to one of the two populations (Figures 2E and 2F), both of which are hallmarks of selection against deleterious mutations (cf., recent findings in humans; Nelson et al. 2012; Fu et al. 2013), and are expected to drive both F^U and F^W downwards, as shown in Materials and Methods (see also Maruki et al. 2012; Bhatia et al. 2013; Jakobsson et al. 2013). For the 4-fold sites on both A and X, no correlation with K_A was observed for F^U, F^W, MAF, and the proportion of private SNPs (Figure 2; $P > 0.1$ in all cases based on Kendall's τ).

The data presented in Figures 1 and 2 suggests that the lack of difference between F_0^W and either F_4^W or F_{SI}^W reported in the previous section is probably due to the fact that F^W gives more weight to SNPs with higher expected levels of polymorphism (e.g., nearly neutral variants), as we have shown in Materials and Methods. In other words, when all 0-fold sites in crossover regions were analysed together (Table 1), the effects of purifying selection on a substantial fraction of 0-fold sites were probably masked by those 0-fold sites that are nearly neutrally evolving. Consequently, the overall distribution of F_0^W appears non-distinguishable from those of F_4^W and F_{SI}^W. In contrast, F^U gives equal weight to all SNPs. Considering that the value of F when calculated using a single SNP is constrained by MAF [see Eq. (3) in Materials and Methods], F^U is expected to be more sensitive to the action of purifying selection than F^W, consistent with the observation reported above. In the Discussion, we will further explore the implications of these statistical properties of F, which arise when information from multiple SNPs is combined.

Longer introns are under stronger selective constraints and are less differentiated

In agreement with earlier findings (Haddrill, Charlesworth, et al. 2005; Halligan and Keightley 2006), longer introns tend to have lower divergence (K) between *D. melanogaster* and *D. simulans* (A: Kendall's $\tau = -0.635$, $P < 0.001$; X: Kendall's $\tau = -0.486$, $P < 0.001$; Figures 3A and 3B), probably as a result of the presence of functional elements that are subject to purifying selection (Bergman and Kreitman 2001; Parsch 2003; Andolfatto 2005; Haddrill, Charlesworth, et al. 2005; Halligan and Keightley

2006; Casillas et al. 2007; Roy et al. 2010). Here, we report further support for this hypothesis by examining within-population polymorphism patterns as a function of intron length. Consistent with the action of purifying selection, longer introns have lower π (Figures 3C and 3D) and more negative Tajima's *D* (Figures 3E and 3F) compared to 4-fold and SI sites (similar results were observed in the FR sample; see Supplementary Figure S7). Interestingly, the patterns of divergence and polymorphism level off for introns longer than 2000 bp. Using the RG sample, the values of π and Tajima's *D* obtained from introns longer than 2000 bp are 0.0072 and -0.5476 for A, and 0.0076 and -0.9013 for X; all these values are substantially lower than the corresponding values observed at 4-fold and SI sites, but are higher than those obtained from 0-fold sites (see Table 1). Furthermore, the K_A values for CDS in crossover regions between *D. melanogaster* and *D. simulans* are 0.015 and 0.018 for A and X, respectively, which are significantly smaller than the values of *K* for long introns > 2000 bp on A and X, which are 0.061 and 0.074, respectively (Mann-Whitney U test, $P < 0.001$). These results imply that long introns, especially those > 2000 bp, are more constrained than the 4-fold and SI sites, but probably contain fewer strongly selected sites than 0-fold sites.

Estimates of F^W, when calculated using sites from introns more than 65 bp in length, were 0.171 and 0.283 for A and for X, respectively. None of these was found to be statistically different from the corresponding values estimating using 4-fold and SI sites reported in Table 1 ($P_{permutation} > 0.1$ in all cases). F^U for introns > 65 bp were 0.157 and 0.174 for A and X, respectively, both of which were significantly smaller than both F_{SI}^U and F_4^U ($P_{permutation} < 0.001$ in all cases). There is a clear negative relationship between F^U and intron length (Figures 4A and 4B; for A and X, Kendall's τ = -0.356 and -0.364; $P = 0.010$ and $P < 0.001$, respectively), which mirrors that between MAF (or the prevalence of private SNPs) and intron length (Supplementary Figure S8), and is consistent with the expected effect of purifying selection on genetic differentiation between populations. The relationship between differentiation and intron length is weaker when F^W was analysed (Supplementary Figure S8; for A and X, Kendall's τ = -0.271 and -0.146, and $P = 0.05$ and 0.16, respectively). These differences between F^W and F^U can be explained by the fact that fewer sites in introns > 65 bp are expected to be strongly selected compared to 0-fold sites. As discussed in the previous section, F^W,

which tends to reflect differentiation patterns at neutral sites in the data, is less likely to recover signatures of purifying selection compared to F^U.

Regions with reduced recombination tend to have higher F

It is known that genomic regions that lack crossing over [non-crossover (NC) regions] have very different patterns of divergence and polymorphism than those seen in crossover (C) regions (Haddrill et al. 2007; Betancourt et al. 2009; Arguello et al. 2010; Campos et al. 2012; Campos et al. 2014). In Table 2, we present summary statistics of the NC data pertinent to the current study (see Materials and Methods for a list of the NC regions considered). It can be seen that, for both 0-fold and 4-fold sites, values of *F* in NC regions are generally higher than those obtained using the same type of site in C regions, regardless of the way in which information from multiple SNPs was combined. Specifically, the average K_A to *D. yakuba* is about 0.05 for the NC loci (Campos et al. 2012). F_0^U calculated using autosomal and X-linked NC data are 0.1817 and 0.3012, respectively (Table 2), higher than the values of 0.1569 and 0.1685 for autosomal and X-linked genes in C regions spanning the same K_A values (Figures 2A and 2B; $P_{permutation} = 0.05$ for A and $P_{permutation} < 0.001$ for X).

It should be noted that the elevation in *F* in NC regions is probably caused by an extreme reduction in within-population diversity induced by tight linkage between a large number of selected sites (Table 2; Kaiser and Charlesworth 2009; O'Fallon et al. 2010; Seger et al. 2010; Zeng and Charlesworth 2010b). This is because *F* is a relative measure of differentiation [see Eq. (1)], and therefore all else being equal, *F* is expected to be elevated by forces that reduce within-population diversity [i.e., π_S in Eq. (1)], irrespective of whether diversifying selection or reduced gene flow has affected the genomic region under study (Charlesworth 1998; Noor and Bennett 2009).

To further examine the effects of selection at linked sites, we inspect the correlation between *F* at putatively neutral sites and local recombination rates in C regions. Figure 5 presents results based on autosomal loci, where it can be seen that F_4^U is reduced with more frequent recombination (Kendall's $\tau = -0.474$, $P = 0.004$; the data point obtained from the NC regions was not included in the calculation). However, there is no statistically significant relationship between recombination rate and F_{SI}^U (Figure 5B; Kendall's $\tau = -0.179$ and $P = 0.28$). Weak negative correlations were also found on the

X chromosome for 4-fold and SI sites (Supplementary Figure S9). The patterns remained unchanged when F^W was used (Supplementary Figure S10).

Discussion

By using the high-quality data provided by the *Drosophila* Population Genomics Project, we have found that evolutionary conserved regions (i.e., genes with lower K_A and longer introns) show clear evidence of more intense on-going purifying selection than less conserved genomic regions, which can be detected by analysing patterns of genetic variation both within and between subpopulations. The negative correlation between π and intron length reported in Figure 3 extends the study by Parsch et al. (2010) who examined a much smaller data set and did not find evidence of such a correlation. Since we did not find support for a correlation between local recombination rate and intron length (Kendall's τ = -0.004 and 0.011 for A and X, respectively, and P > 0.1 in both cases; Supplementary Figure S11) (c.f., Carvalho and Clark 1999; Comeron and Kreitman 2000), the relationship is unlikely to be driven by the well-known positive correlation between diversity and recombination. It is unclear why the effect of intron length levels off for introns longer than 2000bp. Analysis of theoretical models (e.g., Ometto et al. 2005) and improved annotation of non-coding functional elements (e.g., Roy et al. 2010) are both needed to solve this problem. Finally, there is evidence that the severe reduction in within-population diversity in non-crossover regions of the genome induced by tight linkage between selected sites has led to elevated F_{ST} values, but there is limited support for this effect in crossover regions.

Purifying selection as a major determinant of population differentiation

Our analysis reveals (*i*) a positive correlation between K_A and F_0 (Figure 2) and (*ii*) a negative correlation between intron length and F calculated using intronic sites (Figure 4). After examining other aspects of polymorphism and differentiation patterns (Figures 1 and 3), we suggest that the observations can be most readily explained by differential intensity of purifying selection acting on different parts of the genome. Similar observations have also been reported in humans (Barreiro et al. 2008; Maruki et al. 2012), suggesting the universal importance of purifying selection as a factor that shapes genetic differentiation between populations.

It should be noted that the above conclusion is not inconsistent with the existence of outlier loci with unusually high *F*, which may have been caused by diversifying selection (Turner et al. 2008; Yukilevich et al. 2010; Kolaczkowski et al. 2011; Fabian et al. 2012; Langley et al. 2012; Pool et al. 2012; Campo et al. 2013). Our analysis intends to detect forces with large-scale effects (there are typically hundreds of genes in each of the bins in our analysis), and is therefore unlikely to respond to processes that have more localised effects in the genome. In fact, it has been suggested that the number of loci contributing to differences between populations may be relatively small (Yukilevich et al. 2010; Fabian et al. 2012). For example, after taking into account the confounding effects of complex demography and correcting for multiple testing, only four loci had strong statistical support for being driven to high levels of differentiation by diversifying selection between North American and African populations of *D. melanogaster* (Yukilevich et al. 2010). Furthermore, in line with the low level of linkage disequilibrium (LD) in the *D. melanogaster* genome (Pool et al. 2012), previous genome scan studies have shown that most candidate variants that show evidence of involvement in local adaptation only affect differentiation patterns in its immediate neighbourhood, typically on the order of the size of a gene (Kolaczkowski et al. 2011; Fabian et al. 2012). Finally, we have focused on protein-coding regions and introns whereas a substantial number of previously-found candidate loci fall within intergenic regions.

A noticeable exception is chromosome 3R, in which the cosmopolitan inversion *In(3R)P* is situated. Multiple studies concerning differentiation between various *D. melanogaster* populations have found that chromosome 3R has a disproportionally large number of candidate loci, especially within the *In(3R)P* region, and that these candidate variants tend to affect differentiation patterns in a larger genomic neighbourhood (Kolaczkowski et al. 2011; Fabian et al. 2012). To further test the robustness of our results, we repeated the analysis leading to Figure 2A by removing all genes on chromosome 3R, and found that the pattern remains unchanged (Supplementary Figure S12). In summary, it is unlikely that highly differentiated regions driven by adaptive changes have made a substantial contribution to our observations.

The relationship between F and recombination

As pointed out previously (Charlesworth 1998; Noor and Bennett 2009), forces that reduce within-population diversity can lead to elevated F_{ST} values in the absence of diversifying selection and restricted gene flow. Hence, in light of the lack of evidence of adaptive evolution in non-crossover regions of the *D. melanogaster* genome (Betancourt et al. 2009; Arguello et al. 2010; Campos et al. 2014), the high *F* values obtained from NC regions is probably a result of the diversity-reducing effect of linkage between selected sites, which is often referred to Hill-Robertson interference or HRI (Hill and Robertson 1966; Comeron et al. 2008; Sella et al. 2009; Charlesworth 2012b; Cutter and Payseur 2013). Within the C regions, although negative correlations between *F* at putatively neutral sites and local recombination rate, as predicted by the HRI theory, were observed (Figures 5, S9 and S10), these patterns are weak and often non-significant. Langley et al. (2012) also reported weak negative correlations between a different measure of genetic differentiation and fine-scale recombination rates estimated from LD patterns, but the relationship was inconsistent between chromosome arms and was sometimes weakly positive when broad-scale recombination rates were used.

The weak association between *F* and recombination in C regions is somewhat surprising given that both π_4 and π_{SI} are clearly positively correlated with local recombination rates in both the RG and FR populations (Supplementary Figure S13). A possible explanation is that, since hitchhiking effects induced by both positive and negative selection can lead to an excess of low-frequency variants at linked neutral sites (Charlesworth et al. 1993; Braverman et al. 1995; Zeng and Charlesworth 2011), the negative correlation between *F* and recombination may be weakened, if rare variants are more common in low-recombination regions, as these variants tend to lower *F* [see Eq. (3)]. Tajima's *D* is somewhat more negative in autosomal C regions with reduced recombination (Supplementary Figure S14), but it is hard to determine to what extent this has contributed to the observations in Figures 5 and S9, especially when noting that NC regions have more negative Tajima's *D* and yet higher F_{ST} values. Further research that takes into account HRI, demography, and statistical properties of estimators of *F* (see below) is needed to clarify the matter.

The importance of sampling strategy regarding using F_{ST} to study population differentiation

As is the case for other definitions of F_{ST} (Maruki et al. 2012; Jakobsson et al. 2013), Weir and Cockerham's F_{ST} can only take a very restricted range of values when MAF is small [$\max(F_{ST}) \leq 2$ MAF; Supplementary Figure S1]. When information is combined across SNPs, the weighted mean F_{ST} (F^W) is likely to be dominated by SNPs that are more polymorphic (i.e., those having a higher expected MAF). This can lead to the masking of signals of purifying selection, as we have shown above. Thus, F^W may be a better choice when the intention is to ascertain the overall level of genetic differentiation. In this case, as long as the data contains a substantial number of putatively neutrally evolving variants, a reasonably accurate estimate can be obtained, even in the presence of sites under strong selective constraints. In contrast, the unweighted mean F_{ST} (F^U) gives equal weight to all SNPs, and is more responsive to the presence of rare variants (e.g., those under purifying selection). These considerations, as well as the recommendations proposed by Bhatia et al. (2013), suggest that care should be exercised when deciding which sampling strategy is most appropriate for the question in hand.

Tables

Table 1. Summary statistics for loci in crossover (C) regions.

		Within population [a]			Between populations [b]		
Chr	Site	Pop. [c]	π	Tajima's D	MAF	F^U	F^W
A	0-fold	RG	0.0012	-0.8397	0.1222	0.1516	0.1709
		FR	0.0010	-0.2586			
	4-fold	RG	0.0154	-0.1069	0.1653	0.1684	0.1743
		FR	0.0119	0.1116			
	SI [d]	RG	0.0145	-0.1380	0.1630	0.1677	0.1766
		FR	0.0113	0.1413			
X	0-fold	RG	0.0012	-1.1907	0.1073	0.1653	0.2924
		FR	0.0005	-0.2293			
	4-fold	RG	0.0166	-0.4679	0.1367	0.1903	0.2879
		FR	0.0068	0.1412			
	SI [d]	RG	0.0160	-0.4561	0.1379	0.2033	0.3173
		FR	0.0061	0.3414			

[a] Summary statistics calculated using data from within a subpopulation for the type of site under consideration.

[b] Summary statistics calculated using data from both subpopulations for the type of site under consideration. The *F*-statistics are defined by Eqs. (5) and (6).

[c] Population of origin; RG, Rwandan; FR, French.

[d] Sites from 8-30bp regions of short introns ≤ 65bp.

Table 2. Summary statistics for loci in non-crossover (NC) regions.

		Within population			Between populations		
Chr	Site	Pop.	π	Tajima's D	MAF	F^U	F^W
A	0-fold	RG	0.00036	-0.6737	0.1152	0.1817	0.2302
		FR	0.00032	-0.7098			
	4-fold	RG	0.00129	-0.5274	0.1208	0.1906	0.2281
		FR	0.00122	-0.5417			
X	0-fold	RG	0.00056	-0.6392	0.1556	0.3012	0.5673
		FR	0.00023	-0.3126			
	4-fold	RG	0.00327	-0.0084	0.1395	0.2323	0.3485
		FR	0.00090	0.2069			

The statistics were obtained in the same way as in Table 1; see Materials and Methods for more details.

Figures

Figure 1. Polymorphism patterns within 17 Rwandan *D. melanogaster* lines for coding sequence (CDS) binned by K_A value (to *D. yakuba*), and for sites in the 8-30 bp regions of short introns ≤ 65 bp (SI sites). (A) Nucleotide diversity (π) for autosomal CDS where crossing-over occurs (CDS-C) and (B) X-linked CDS-C regions; (C) Tajima's *D* for autosomal CDS-C regions and (D) X-linked CDS-C regions. The x-axes show the maximum K_A value in each bin. Symbols: 0-fold degenerate sites—open circles; 4-fold degenerate sites—open triangles; SI sites—open red squares.

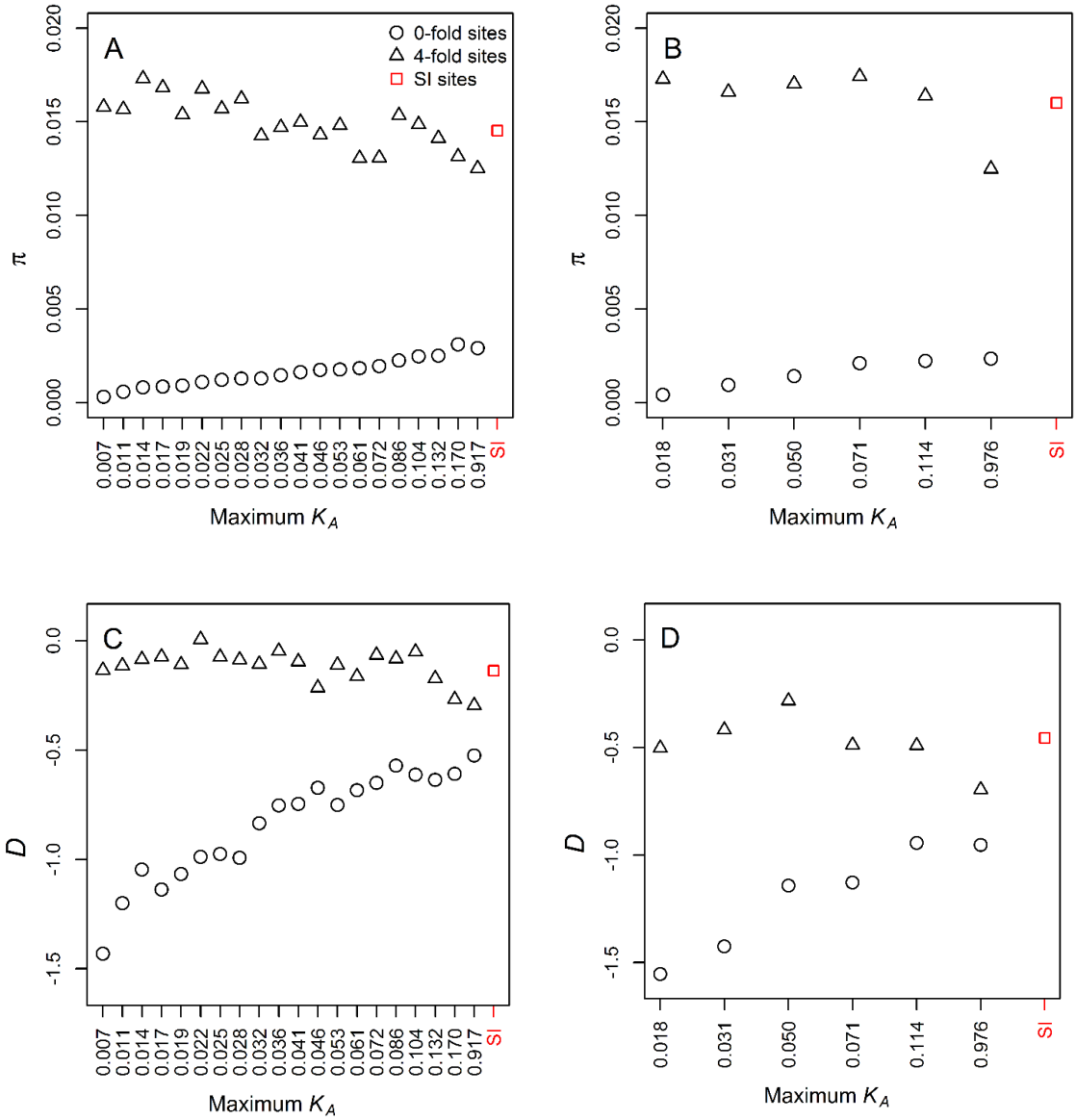
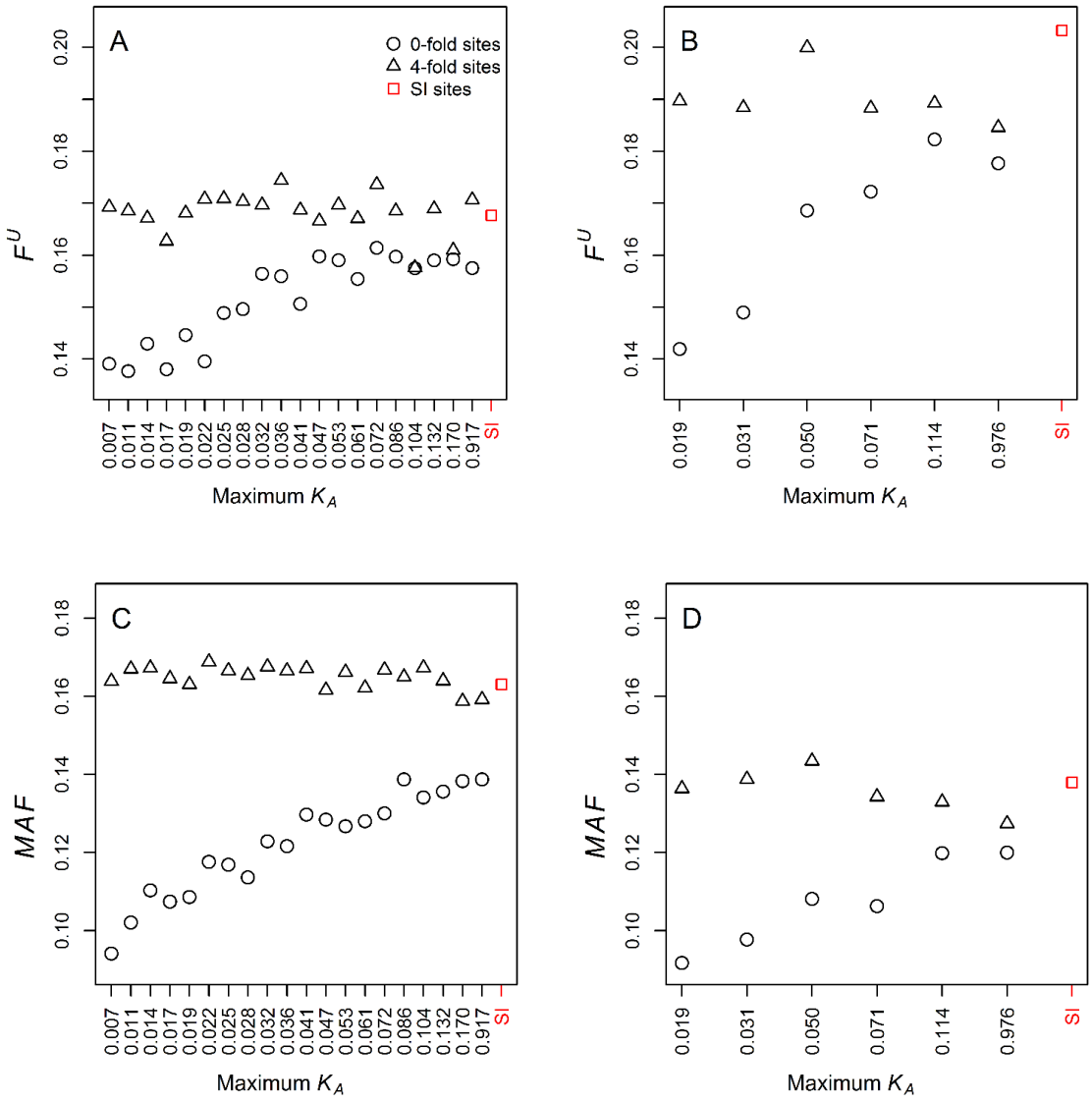
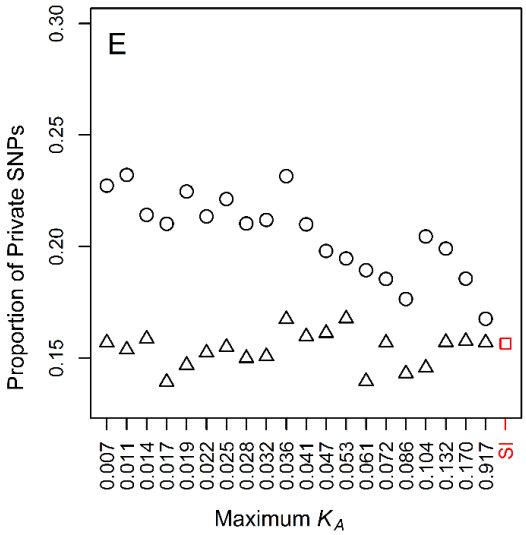


Figure 2. Differentiation patterns between 7 French and 17 Rwandan *D. melanogaster* lines for coding sequence (CDS) binned by K_A value (to *D. yakuba*), and for SI sites. (A) Unweighted mean F_{ST} [F^U; Eq. (5)] for autosomal coding CDS where crossing-over occurs (CDS-C) and (B) X-linked CDS-C regions; (C) population-average minor allele frequency (MAF) for autosomal CDS-C regions and (D) X-linked CDS-C regions; (E) the proportion of SNPs per bin in which one allele was private to one of the *D. melanogaster* populations for autosomal CDS-C regions and (F) X-linked CDS-C regions. Symbols: 0-fold degenerate sites—open circles; 4-fold degenerate sites—open triangles; SI sites—open red squares.




E
Proportion of Private SNPs
0.15
0.20
0.25
0.30
0.007
0.011
0.014
0.017
0.019
0.022
0.025
0.028
0.032
0.036
0.041
0.047
0.053
0.061
0.072
0.086
0.104
0.132
0.170
0.917
SI
Maximum K_A

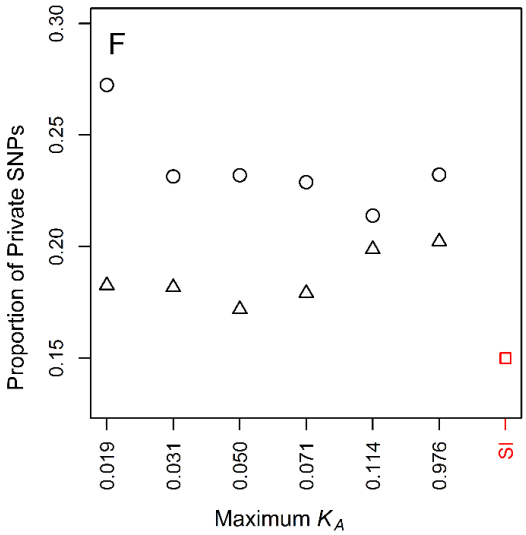
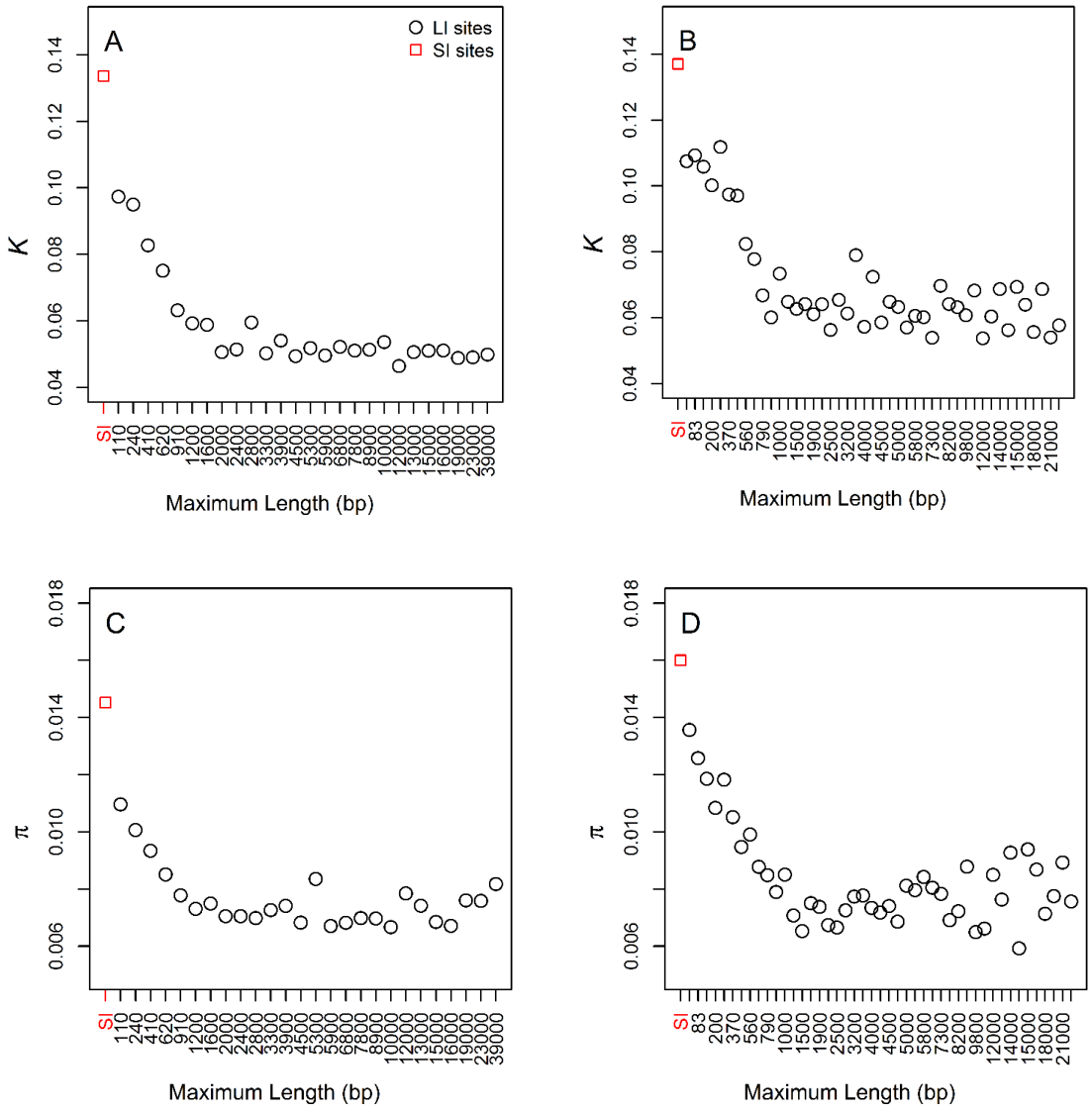
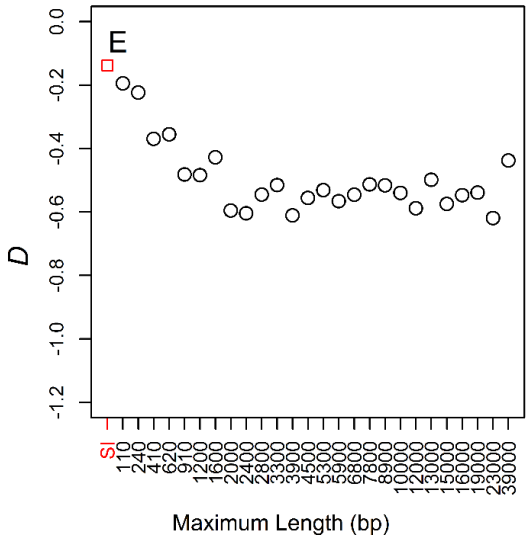

F
Proportion of Private SNPs
0.15
0.20
0.25
0.30
0.019
0.031
0.050
0.071
0.114
0.976
SI
Maximum K_A

Figure 3. Divergence and polymorphism patterns for intronic sites binned by intron length. (A) Divergence (*K*) between *D. melanogaster* and *D. simulans* for autosomal introns and (B) X-linked introns; (C) nucleotide diversity (π) for autosomal introns and (D) X-linked introns; (E) Tajima's *D* for autosomal introns and (F) X-linked introns. The x-axes display the maximum intron length in each bin. Note that the number of SNPs in each autosomal intron bin is roughly the same as that in the autosomal SI bin; the same applies to the X-linked data. Symbols: Long intronic sites—open circles; positions 8-30bp sites of short introns ≤ 65 bp (SI sites)—open red squares.




E
D
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
SI
110
240
410
620
910
1200
1600
2000
2400
2800
3300
3900
4500
5300
5900
6800
7800
8900
10000
12000
13000
15000
16000
19000
23000
39000
Maximum Length (bp)

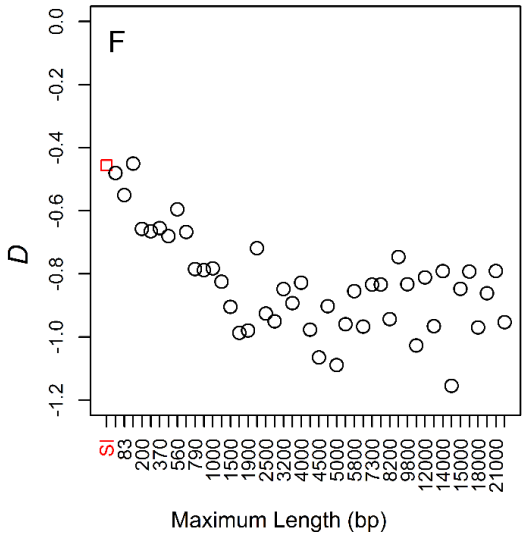

F
D
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
SI
83
200
370
560
790
1000
1500
1900
2500
3200
4000
4500
5000
5800
7300
8200
9800
12000
14000
15000
18000
21000
Maximum Length (bp)

Figure 4. Differentiation between 7 French and 17 Rwandan *D. melanogaster* lines for long intronic sites binned by intron length. (A) Unweighted mean F_{ST} [F^U; Eq. (5)] for autosomal introns and (B) X-linked introns. Symbols: Long intronic sites—open circles; SI sites—open red squares.

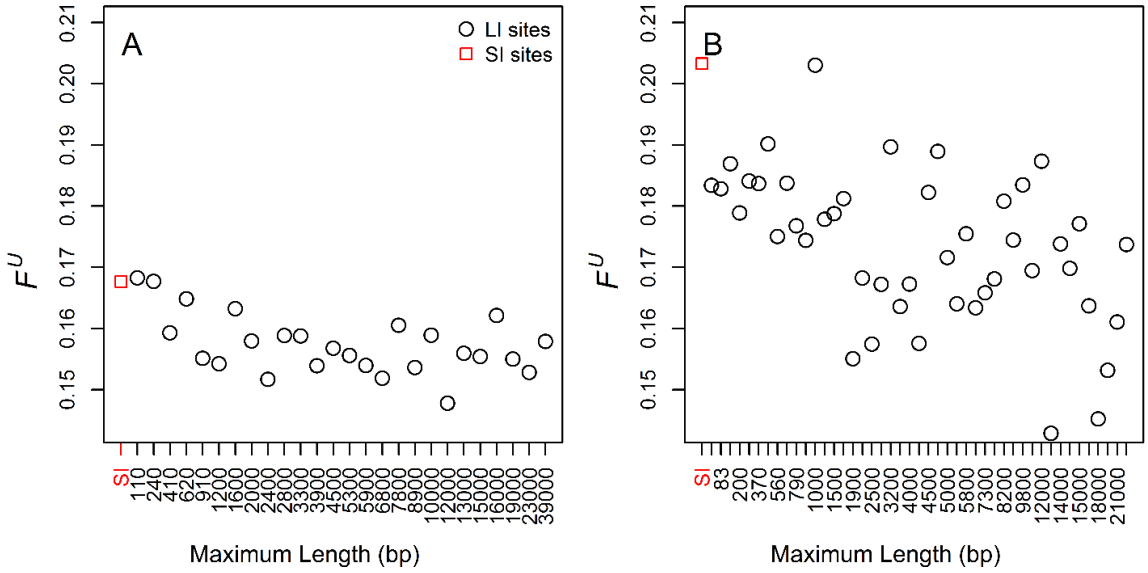


Figure 5. Differentiation between 7 French and 17 Rwandan *D. melanogaster* lines for 4-fold degenerate sites and SI sites in crossover regions as a function of local recombination rate. (A) F^U for autosomal CDS regions and (B) autosomal SI regions.

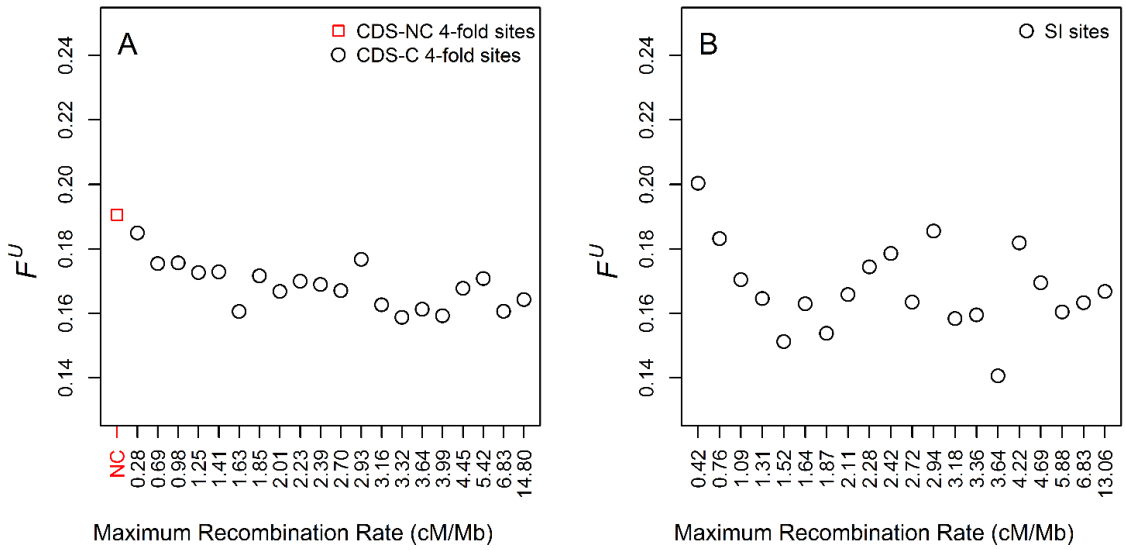


Figure S1. Differences between max(*F*) derived under two different definitions of *F*. The solid line shows the upper bound of *F* as a function of MAF derived in this study using the definition of *F* proposed by Weir and Cockerham (1984). The dotted line shows the upper bound derived by other investigators (Maruki et al. 2012; Jakobsson et al. 2013) using Nei's definition (Nei 1973).

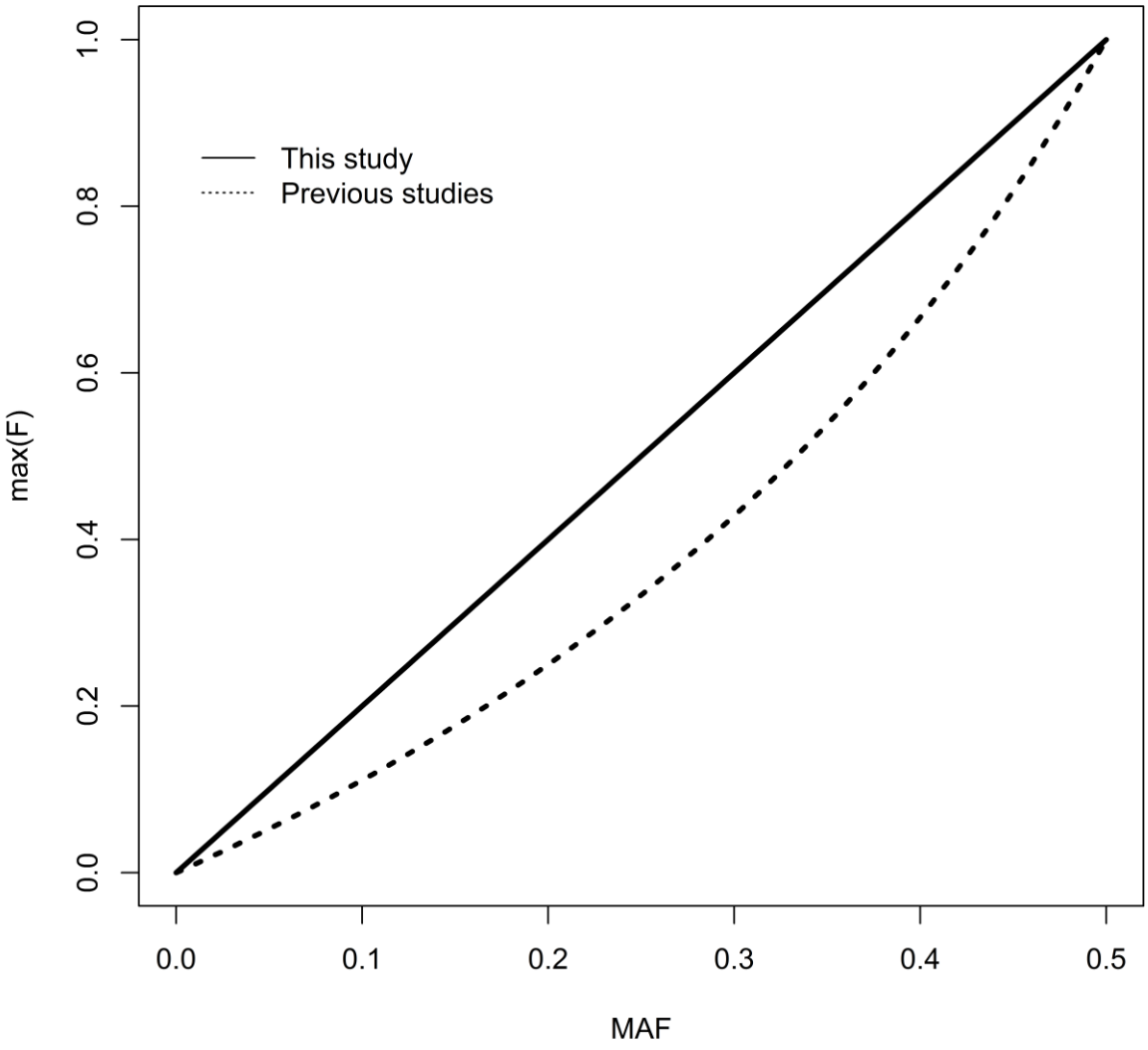
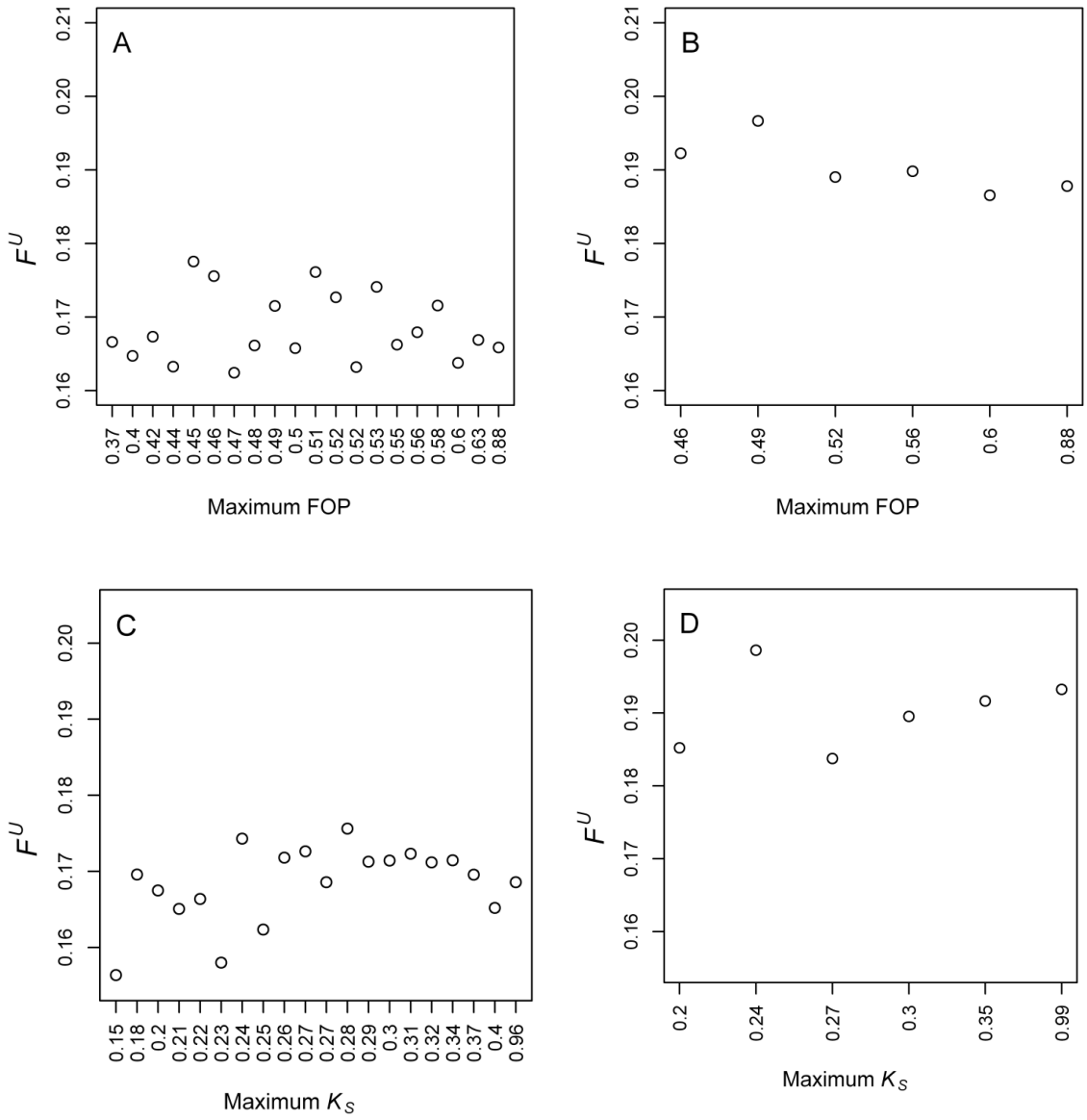
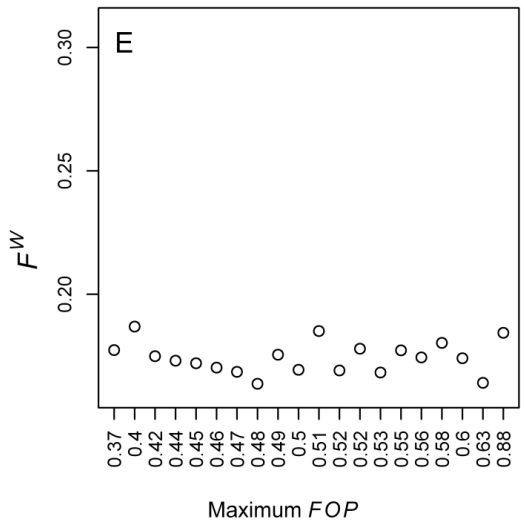
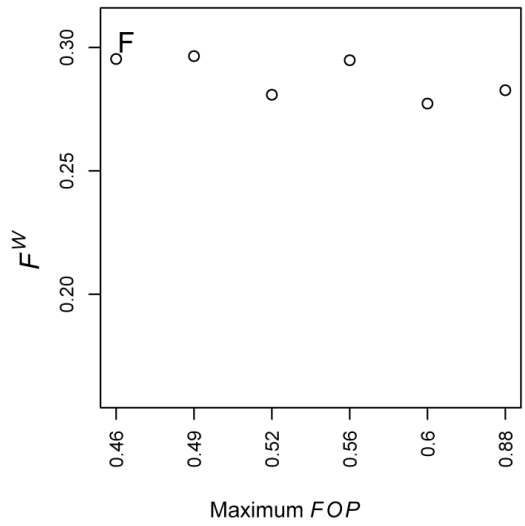
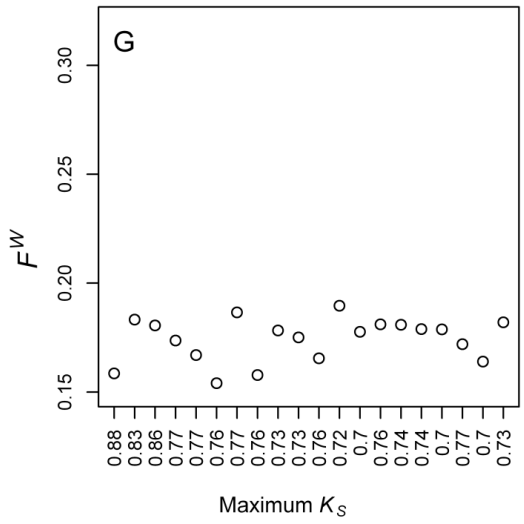


Figure S2. Differentiation between 7 French and 17 Rwandan *Drosophila melanogaster* lines at 4-fold degenerate sites in regions where crossing over occurs, binned by proxies of codon usage bias. (A) Unweighted mean F_{ST} (F^U) binned by the frequency of optimal codons (*FOP*) for autosomal sites, and (B) for X-linked sites; (C) F^U binned by K_S (to *D. yakuba*) for autosomal sites, and (D) for X-linked sites; (E) weighted average F_{ST} (F^W) binned by *FOP* for autosomal sites, and (F) for X-linked sites; (G) F^W binned by K_S (to *D. yakuba*) for autosomal sites, and (H) for X-linked sites.




E
F^W
0.20
0.25
0.30
0.37 0.4 0.42 0.44 0.45 0.46 0.47 0.48 0.49 0.5 0.51 0.52 0.52 0.53 0.55 0.56 0.58 0.6 0.63 0.88
Maximum FOP


F
F^W
0.20
0.25
0.30
0.46 0.49 0.52 0.56 0.6 0.88
Maximum FOP


G
F^W
0.15
0.20
0.25
0.30
0.88 0.83 0.86 0.77 0.77 0.76 0.77 0.76 0.73 0.73 0.76 0.72 0.7 0.76 0.74 0.74 0.7 0.77 0.7 0.73
Maximum K_S

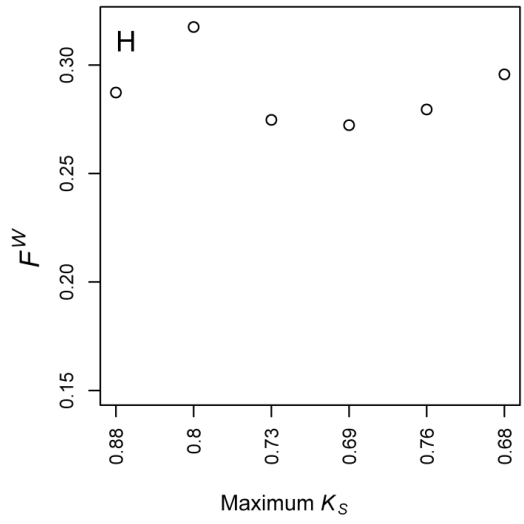

H
F^W
0.15
0.20
0.25
0.30
0.88 0.8 0.73 0.69 0.76 0.68
Maximum K_S

Figure S3. Polymorphism within 7 French *D. melanogaster* lines, for coding sequence (CDS) binned by K_A value (to *D. yakuba*), and for positions 8-30bp of introns ≤ 65 bp in length (SI sites). (A) Nucleotide diversity (π) for autosomal CDS where crossing-over occurs (CDS-C) and (B) X-linked CDS-C regions; (C) Tajima's *D* for autosomal CDS-C regions and (D) X-linked CDS-C regions. Symbols: 0-fold degenerate sites—open circles; 4-fold degenerate sites—open triangles; SI sites—open red squares.

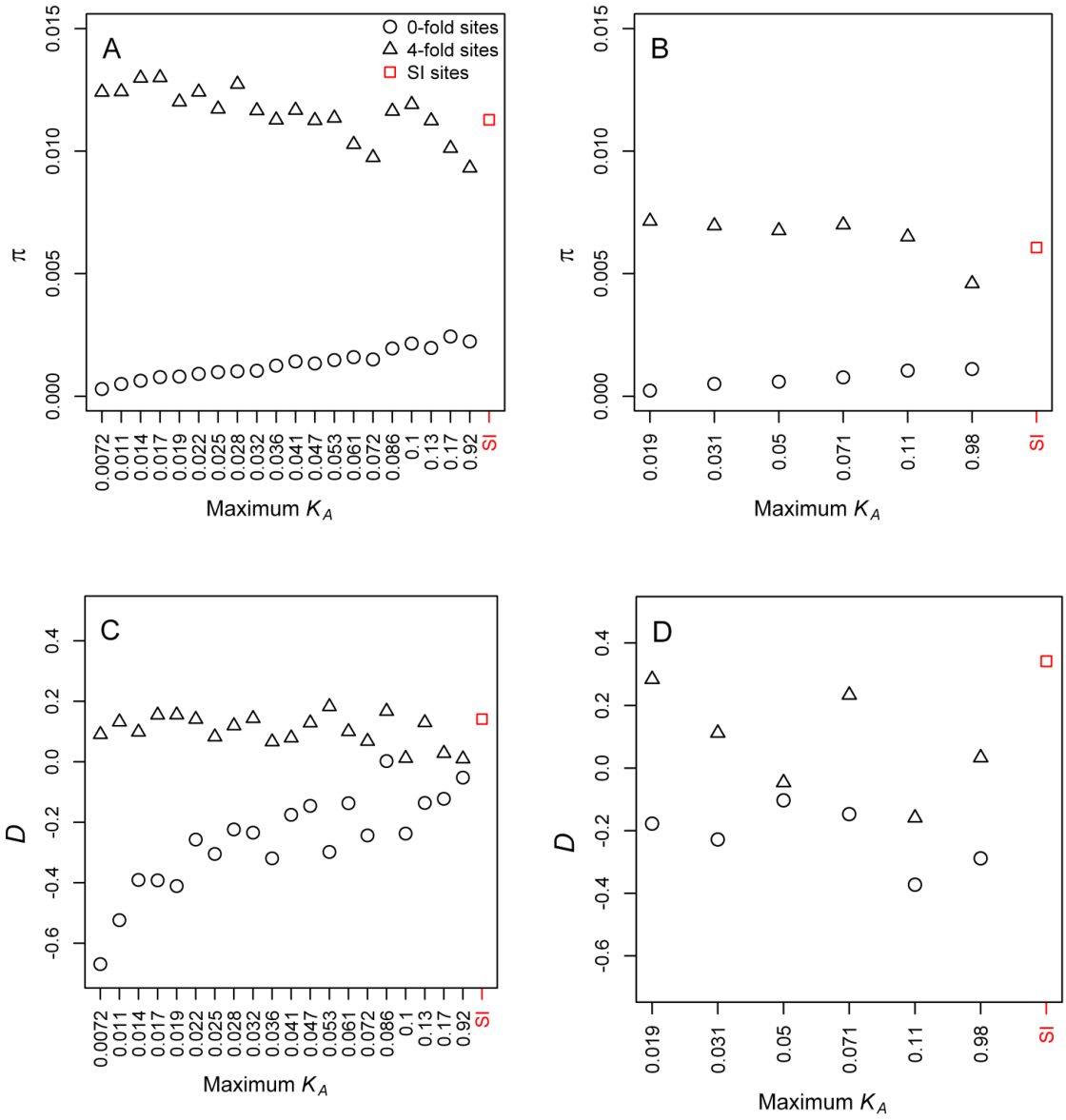


Figure S4. Polymorphism patterns within 17 Rwandan *D. melanogaster* lines for coding sequence (CDS) binned by K_A value (to *D. simulans*), and for sites in the 8-30 bp regions of short introns ≤ 65 bp (SI sites). (A) Nucleotide diversity (π) for autosomal CDS where crossing-over occurs (CDS-C) and (B) X-linked CDS-C regions; (C) Tajima's *D* for autosomal CDS-C regions and (D) X-linked CDS-C regions. The x-axes show the maximum K_A value in each bin. Symbols: 0-fold degenerate sites—open circles; 4-fold degenerate sites—open triangles; SI sites—open red squares.

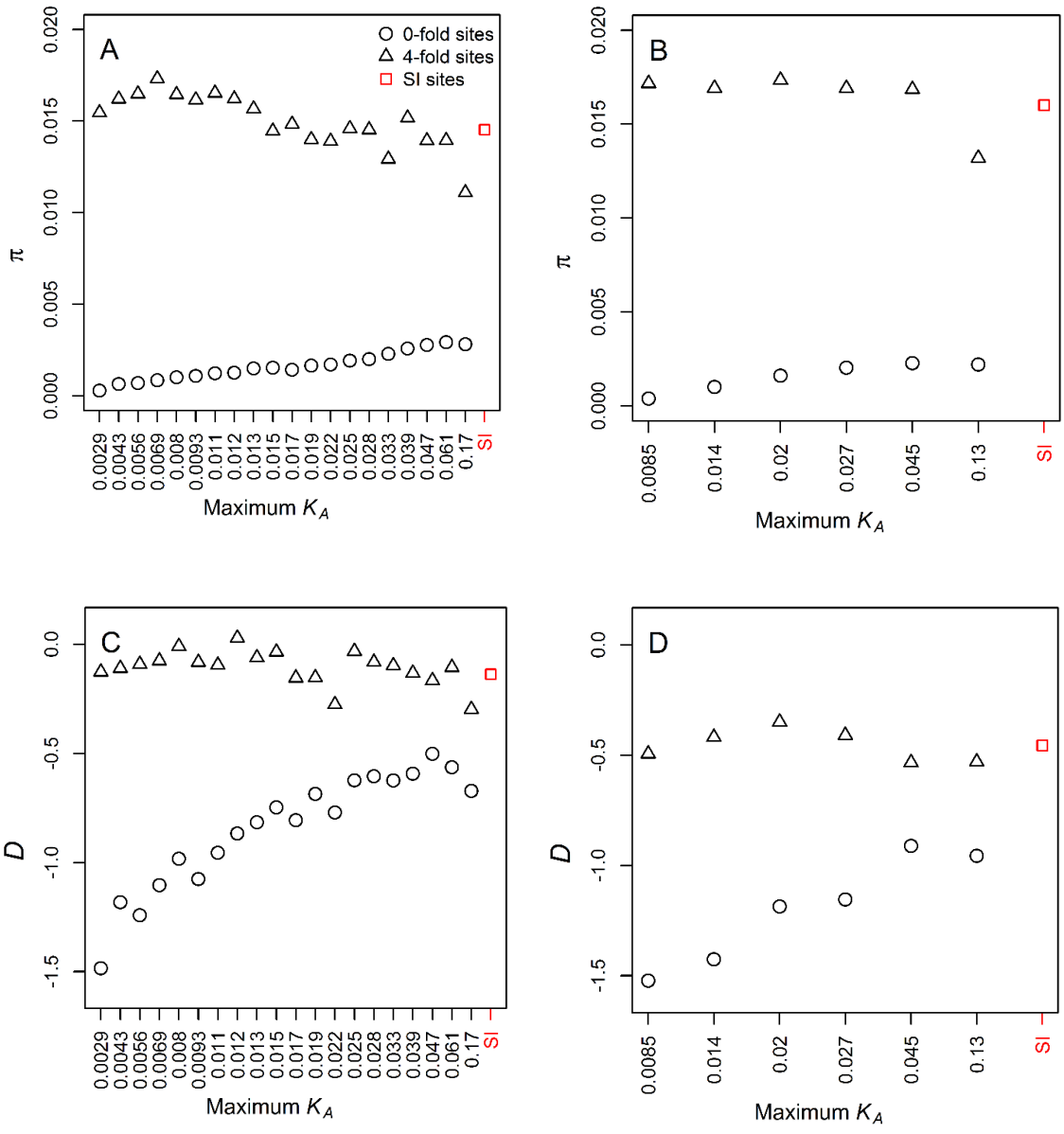
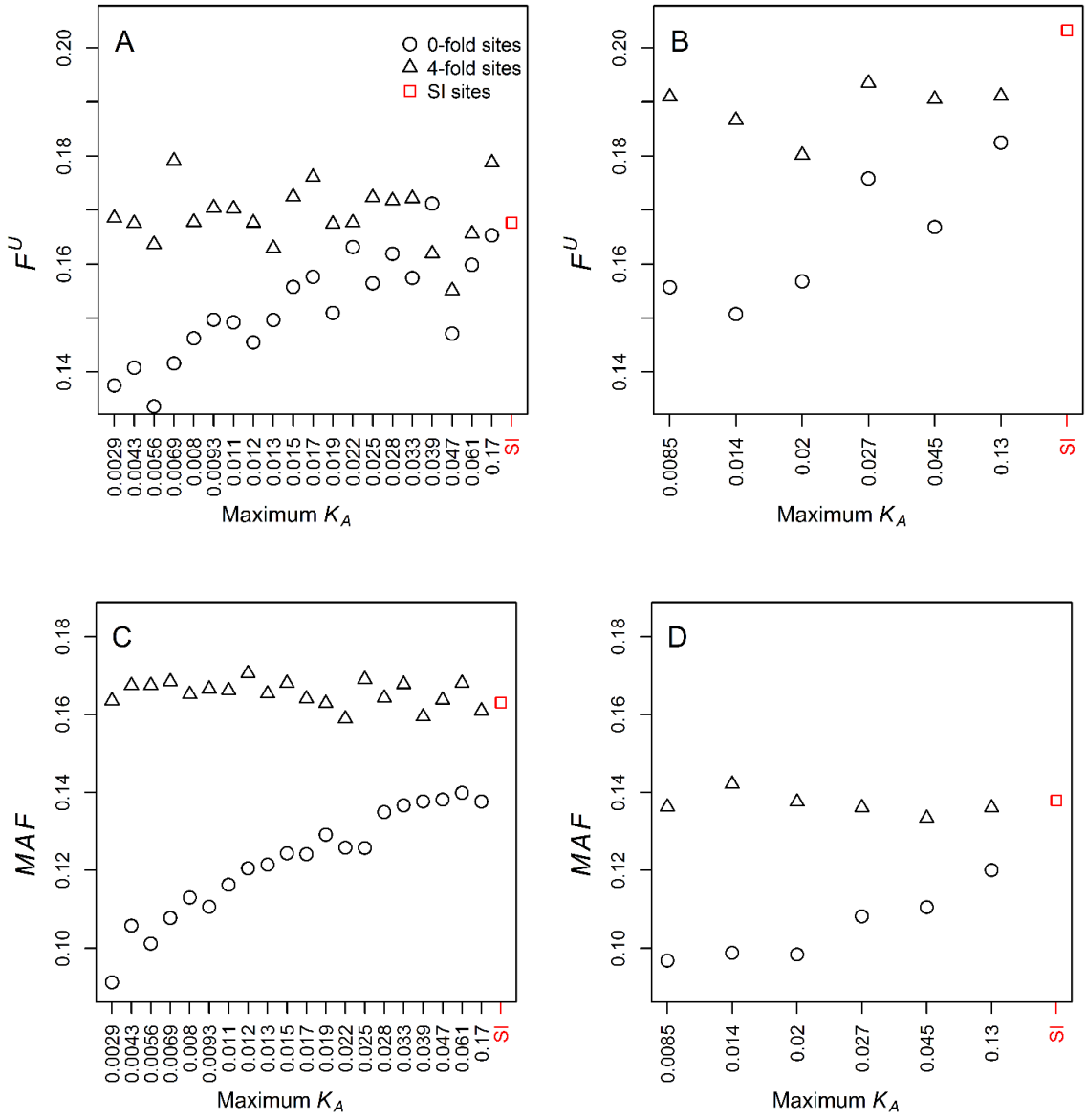


Figure S5. Differentiation patterns between 7 French and 17 Rwandan *D. melanogaster* lines for coding sequence (CDS) binned by K_A value (to *D. simulans*), and for SI sites. (A) Unweighted mean F_{ST} [F^U; Eq. (5)] for autosomal coding CDS where crossing-over occurs (CDS-C) and (B) X-linked CDS-C regions; (C) population-average minor allele frequency (MAF) for autosomal CDS-C regions and (D) X-linked CDS-C regions; (E) the proportion of SNPs per bin in which one allele was private to one of the *D. melanogaster* populations for autosomal CDS-C regions and (F) X-linked CDS-C regions. Symbols: 0-fold degenerate sites—open circles; 4-fold degenerate sites—open triangles; SI sites—open red squares.



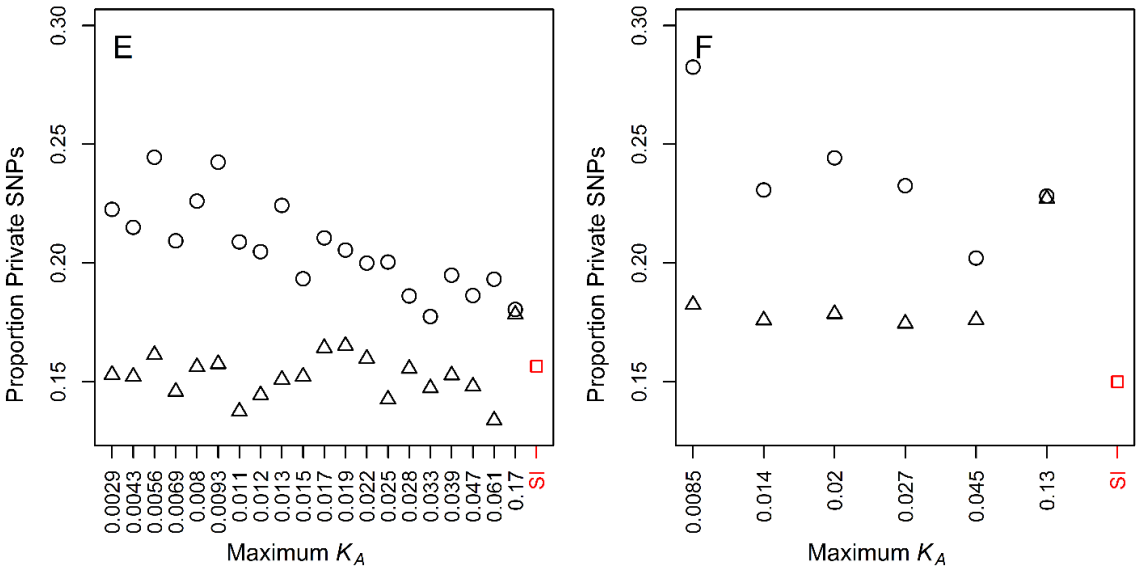

E
F
Proportion Private SNPs
0.15
0.20
0.25
0.30
0.0029
0.0043
0.0056
0.0069
0.008
0.0093
0.011
0.012
0.013
0.015
0.017
0.019
0.022
0.025
0.028
0.033
0.039
0.047
0.061
0.17
SI
0.0085
0.014
0.02
0.027
0.045
0.13
Maximum K_A

Figure S6. Differentiation between 7 French and 17 Rwandan *D. melanogaster* lines, for coding sequence (CDS) binned by K_A value (to *D. yakuba*), and for positions 8-30bp of introns ≤ 65bp in length (SI sites). (A) Weighted mean F_{ST} (F^W) for 0-fold degenerate sites in autosomal CDS where crossing-over occurs (CDS-C) (Kendall's $\tau = 0.4$, P = 0.015) and (B) 4-fold degenerate and SI sites in autosomal CDS-C (Kendall's $\tau = 0.0947$, P = 0.58, data not including the SI point); (C) F^W for 0-fold degenerate sites in X-linked CDS-C (Kendall's $\tau = 0.467$, P = 0.26 and (D) 4-fold degenerate and SI sites in X-linked CDS-C (Kendall's $\tau = 0.467$, P = 0.26. Symbols: 0-fold degenerate sites—open circles; 4-fold degenerate sites—open triangles; SI sites—open red squares.

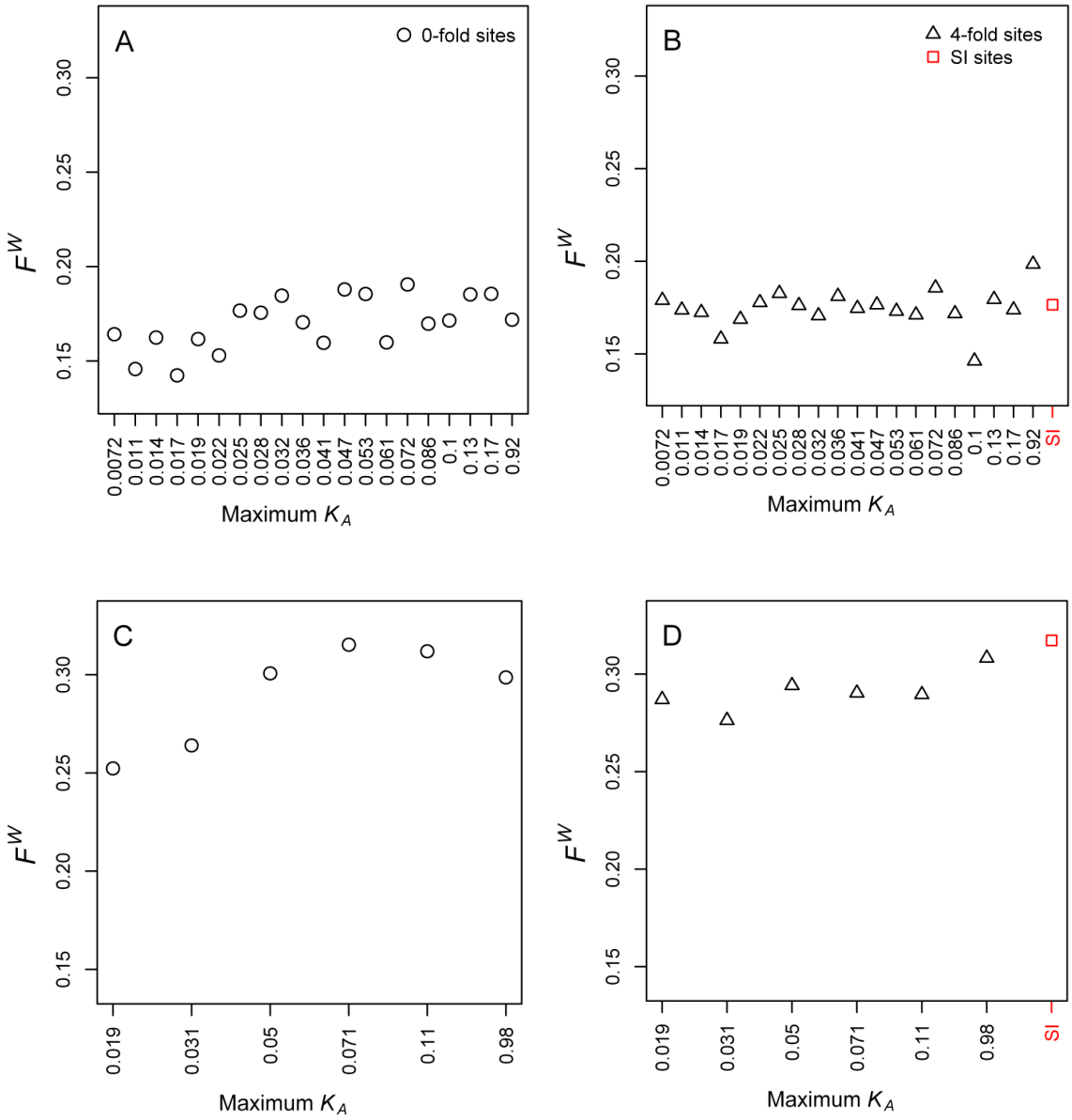


Figure S7. Polymorphism within 7 French *Drosophila melanogaster* lines, for intronic sites > 65bp (long introns), binned by *K* value (to *D. simulans*), with comparison to positions 8-30bp of introns ≤ 65bp in length (SI sites). (A) Nucleotide diversity (π) for autosomal introns and (B) X-linked introns; (C) Tajima's *D* for autosomal introns and (D) X-linked introns. Symbols: long introns sites—open circles; SI sites—open red squares.

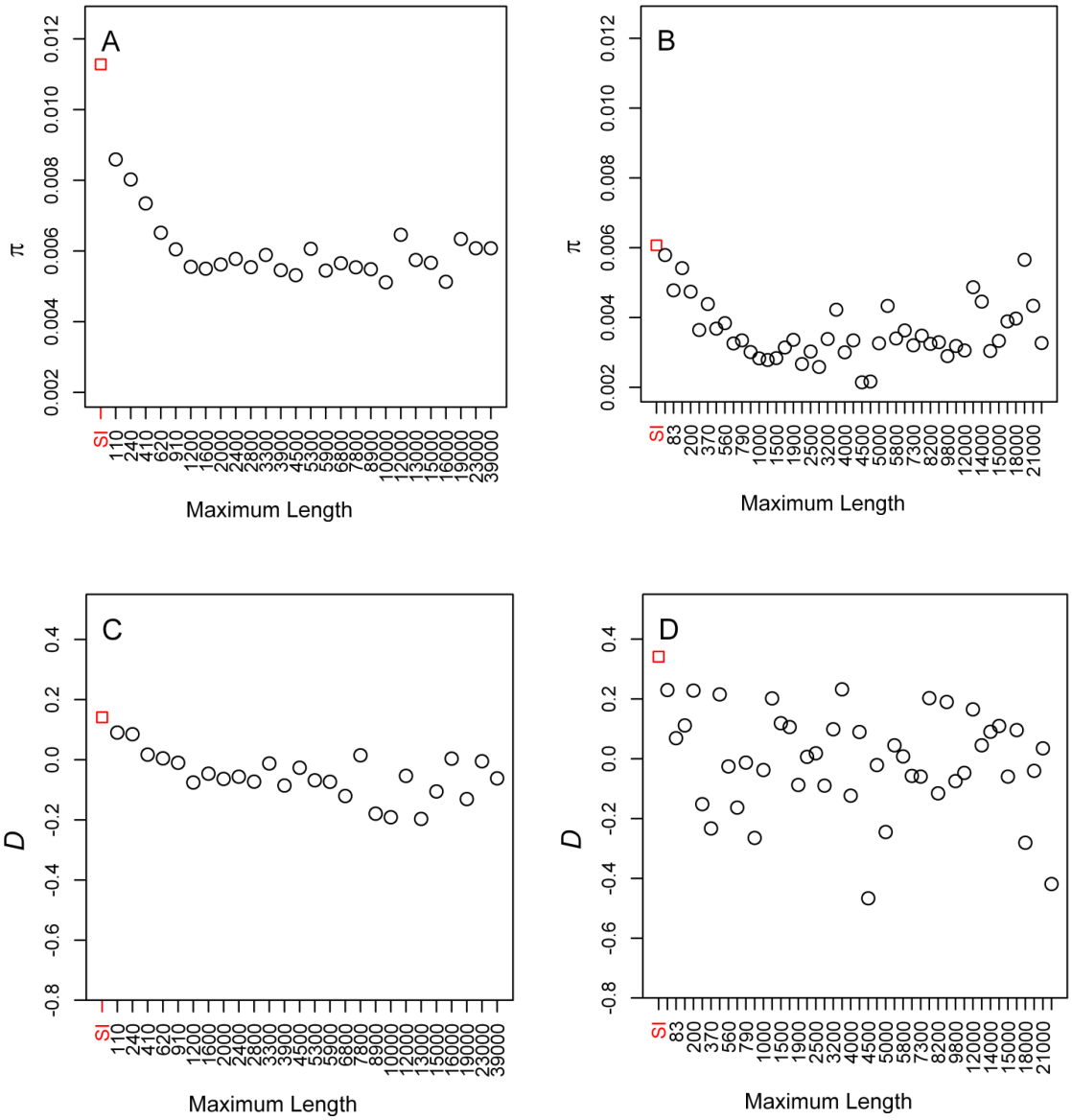
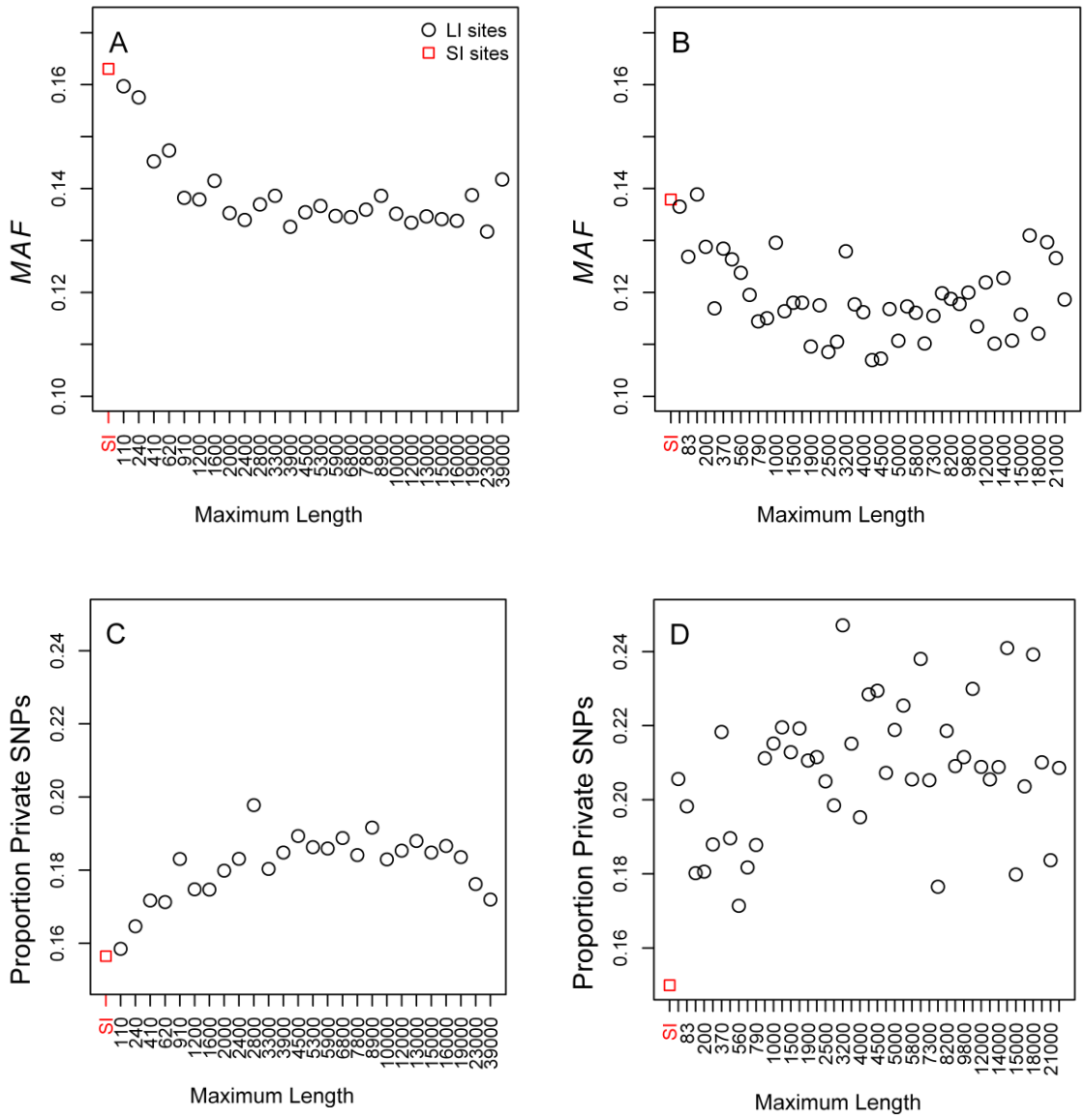
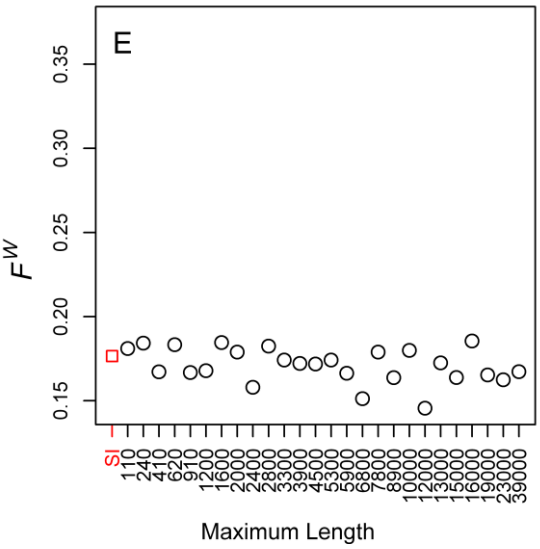


Figure S8. Differentiation between 7 French and 17 Rwandan *Drosophila melanogaster* lines, for intronic sites > 65bp (long introns), binned by intron length, with comparison to positions 8-30bp of introns ≤ 65bp in length (SI sites). (A) Population-average minor allele frequency (*MAF*) for autosomal introns and (B) X-linked introns; (C) proportion of private alleles for autosomal introns and (D) X-linked introns; (E) weighted mean F_{ST} (F^{W}) for autosomal introns and (F) X-linked introns. Symbols: long intron sites—open circles; SI sites—open red squares.




E
F^W
0.15
0.20
0.25
0.30
0.35
SI
110
240
410
620
910
1200
1600
2000
2400
2800
3300
3900
4500
5300
5900
6800
7800
8900
10000
12000
13000
15000
16000
19000
23000
39000
Maximum Length

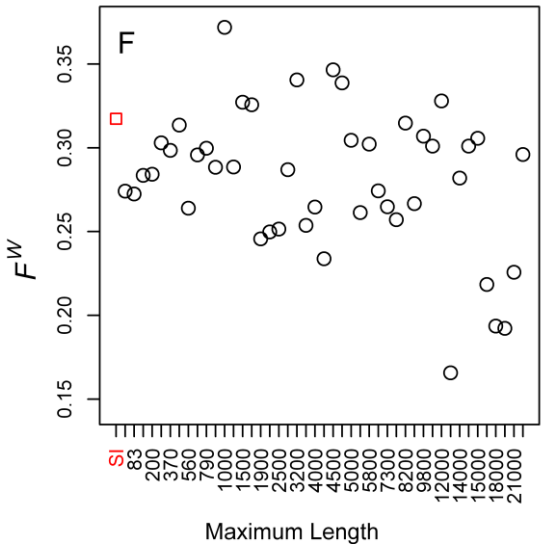

F
F^W
0.15
0.20
0.25
0.30
0.35
SI
83
200
370
560
790
1000
1500
1900
2500
3200
4000
4500
5000
5800
7300
8200
9800
12000
14000
15000
18000
21000
Maximum Length

Figure S9. Differentiation between 7 French and 17 Rwandan *Drosophila melanogaster* lines for X-linked 4-fold degenerate sites and positions 8-30bp of introns ≤ 65bp in length (SI sites), as a function of local recombination rate. Unweighted mean F_{ST} (F^U) for X-linked 4-fold degenerate sites (A) and X-linked SI sites (B), respectively. The red square in (A) represents data from 4-fold sites in NC regions. Kendall's τ for the two figures are -0.2 (P = 0.71) and -0.7 (P = 0.06), respectively

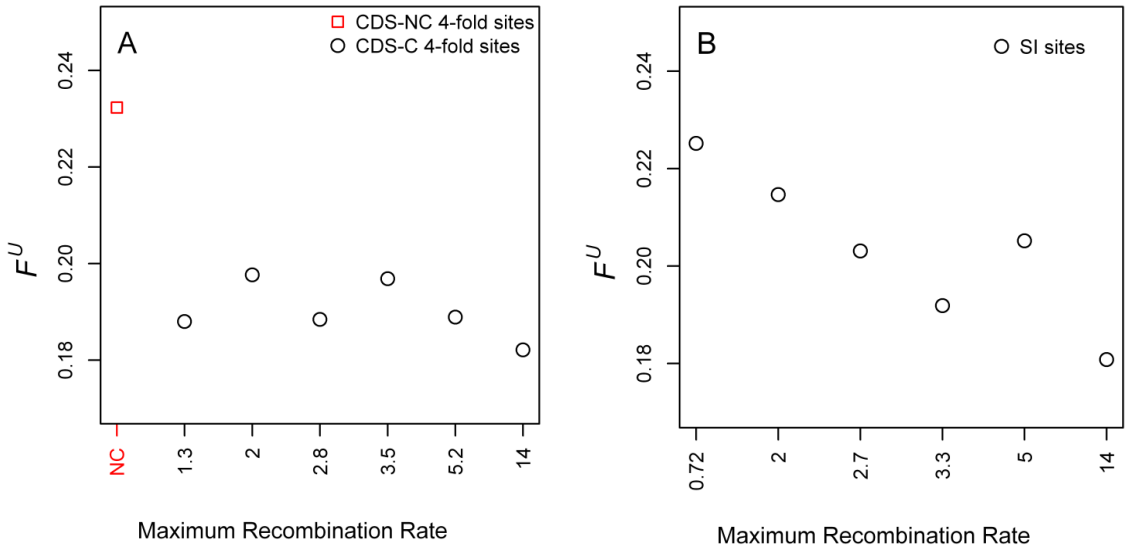


Figure S10. Differentiation between 7 French and 17 Rwandan *Drosophila melanogaster* lines, for 4-fold degenerate sites in coding sequence where crossing over occurs (CDS-C), as a function of local recombination rate, and for coding sequence where crossing over does not occur (CDS-NC). (A) Weighted mean F_{ST} (F^W) for autosomal CDS (Kendall's τ = -0.337, P = 0.041, data not including the NC point), (B) X-linked CDS (Kendall's τ = -0.467, P = 0.26), (C) autosomal SI sites (Kendall's τ = -0.0526, P = 0.77), and (D) X-linked SI sites (Kendall's τ = -0.6, P = 0.13).

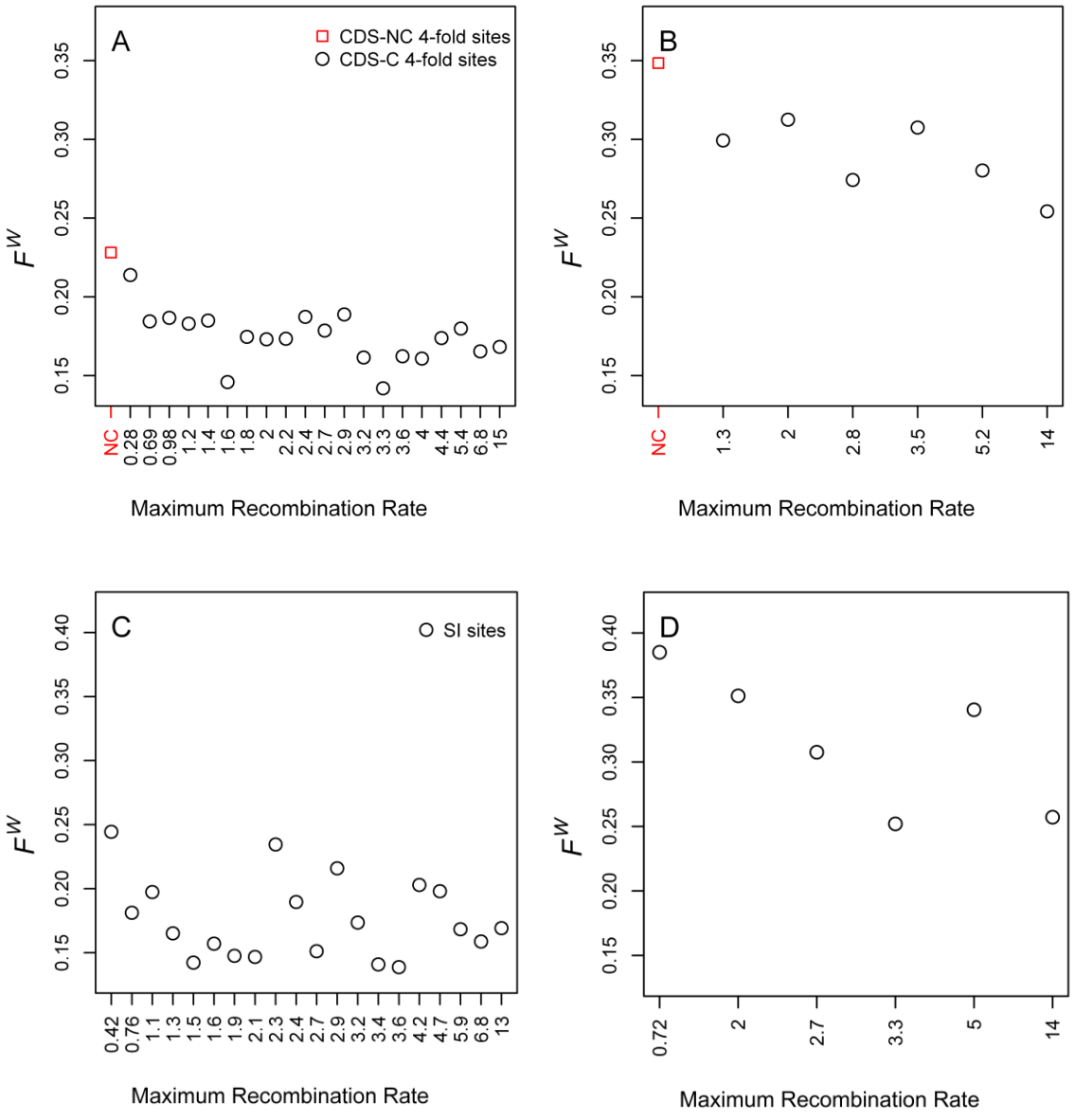


Figure S11. Correlation between intron length and local recombination rate.
Autosomal and X-linked introns are shown in (A) and (B), respectively. The x-axes give the maximum local recombination rates for the recombination rate bins.

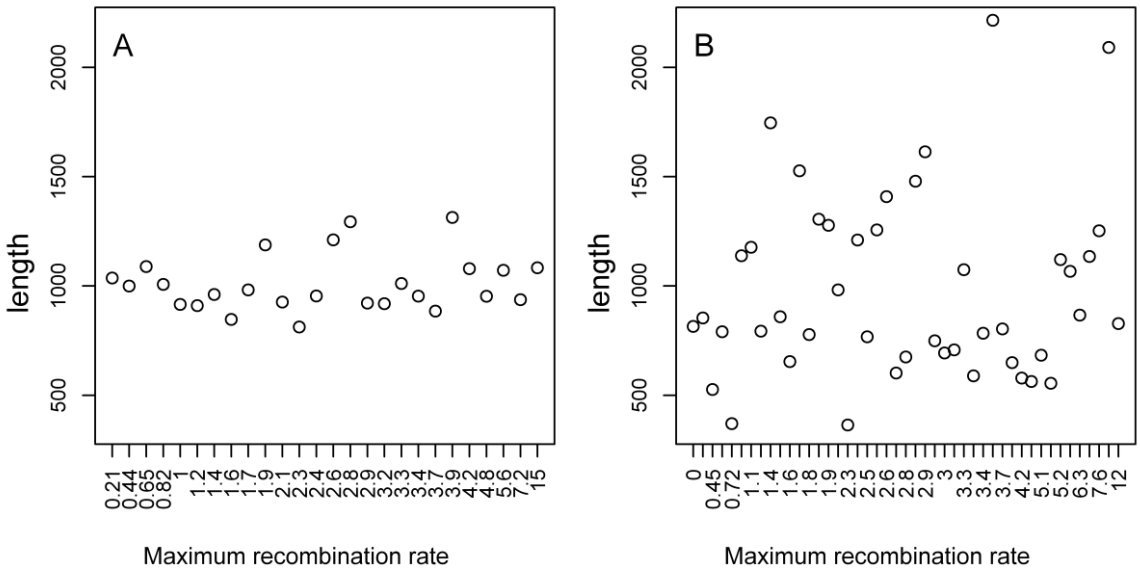


Figure S12. Unweighted mean F_{ST} (F^U) between 7 French and 17 Rwandan *Drosophila melanogaster* lines for autosomal coding sequence where crossing over occurs, and after removing data from chromosome 3R, binned by K_A (to *D. yakuba*). Symbols: 0-fold degenerate sites—open circles; 4-fold degenerate sites—open triangles; SI sites—open red squares.

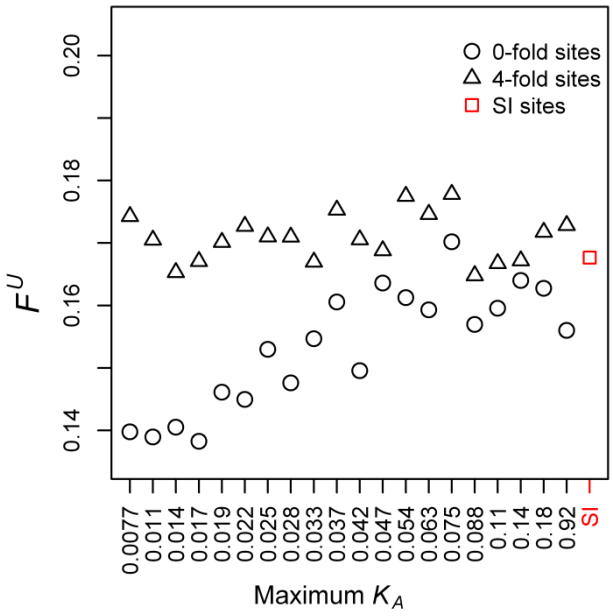
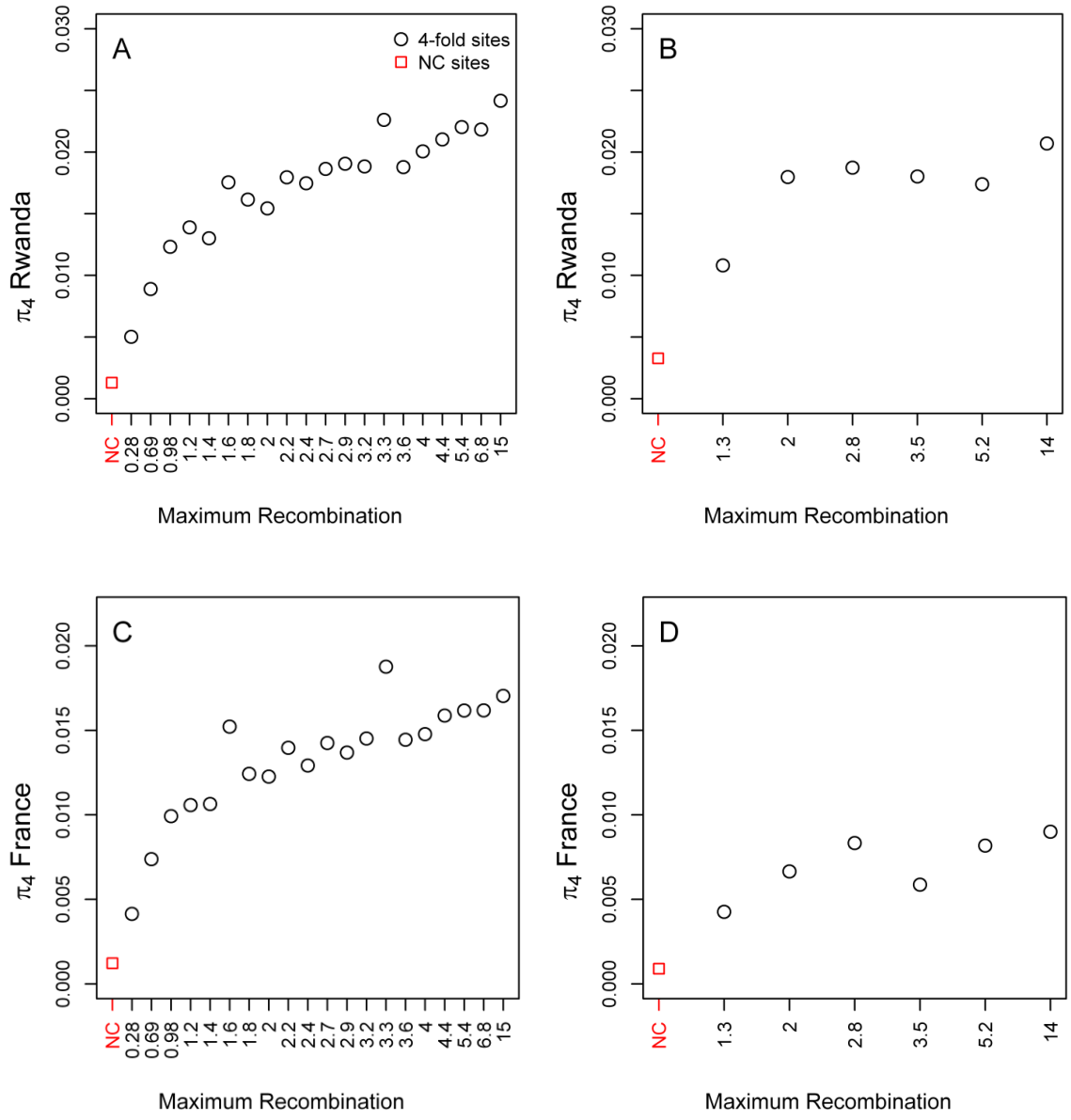


Figure S13. Nucleotide diversity (π) at 4-fold degenerate sites and positions 8-30bp of introns ≤ 65bp in length (SI sites), as a function of local recombination rate. (A) π at autosomal 4-fold degenerate sites in regions where crossing over occurs (CDS-C) and in regions where crossing over does not occur (CDS-NC), Rwandan sample; (B) π at X-linked 4-fold CDS-C and CDS-NC, Rwandan sample; (C) π at autosomal 4-fold CDS-C and CDS-NC, French sample; (D) π at X-linked 4-fold CDS-C and CDS-NC, French sample; (E) π at autosomal SI sites, Rwandan sample; (F) π at X-linked SI sites, Rwandan sample; (G) π at autosomal SI sites, French sample; (H) π at X-linked SI sites, French sample.



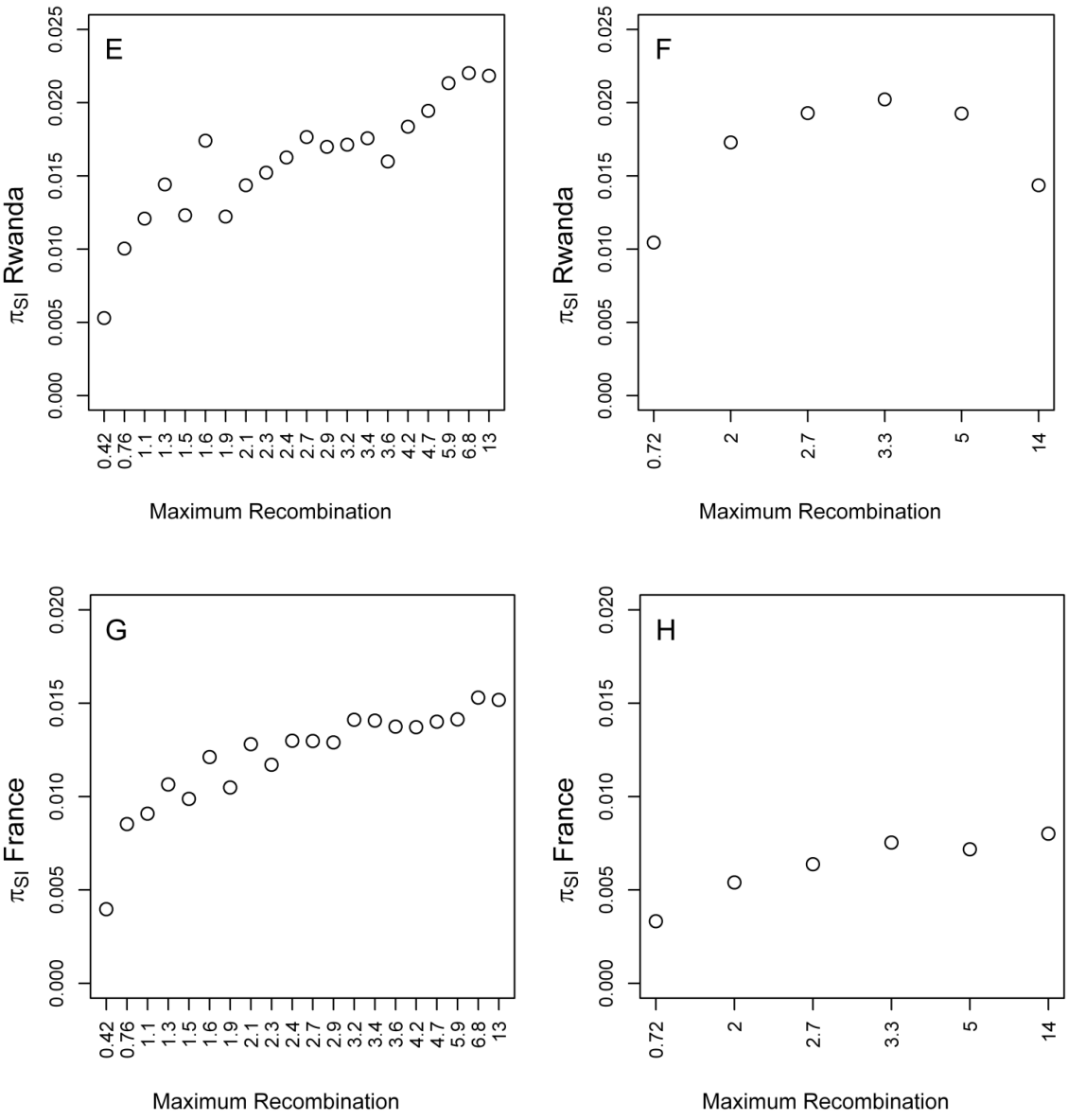
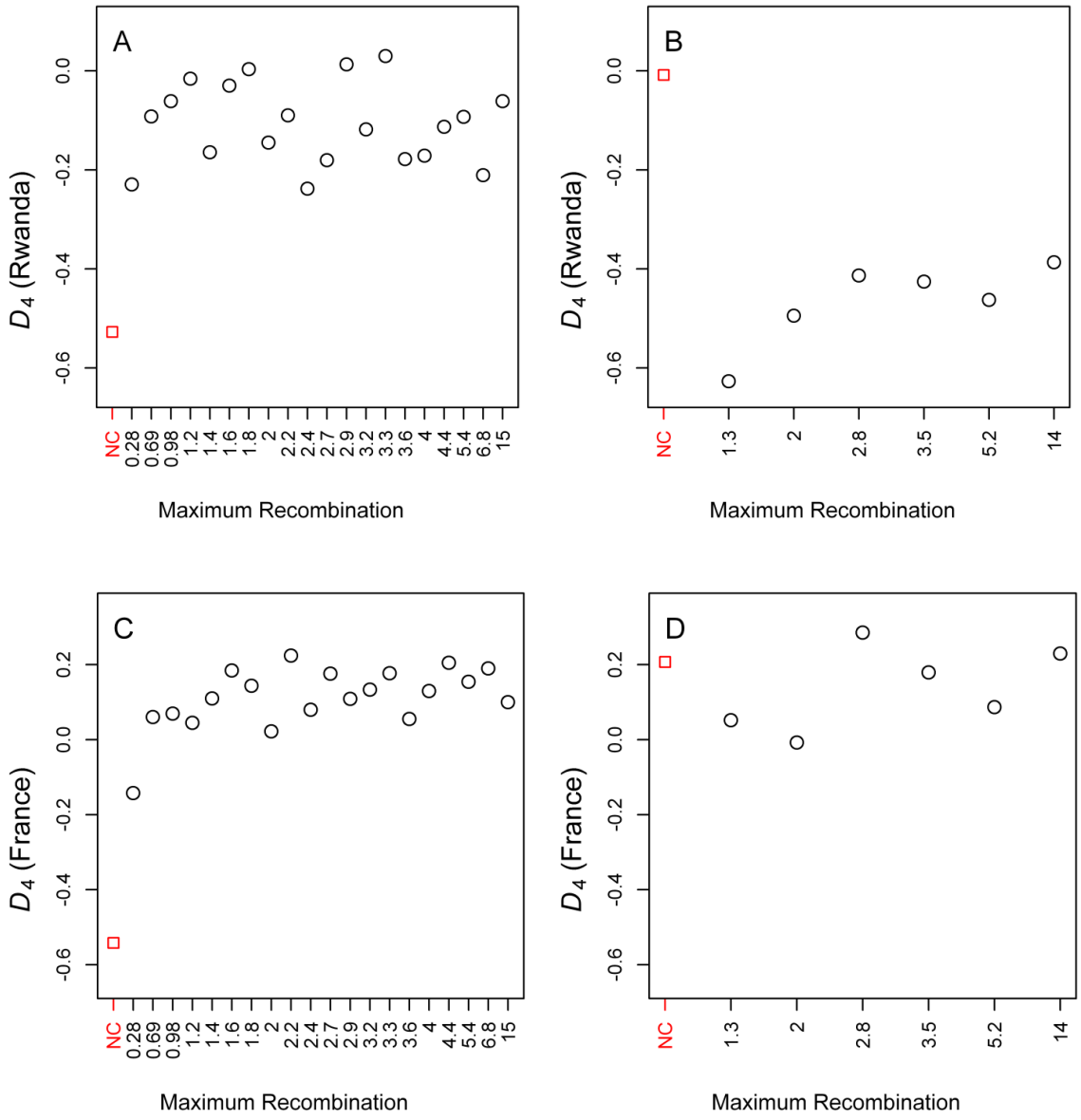
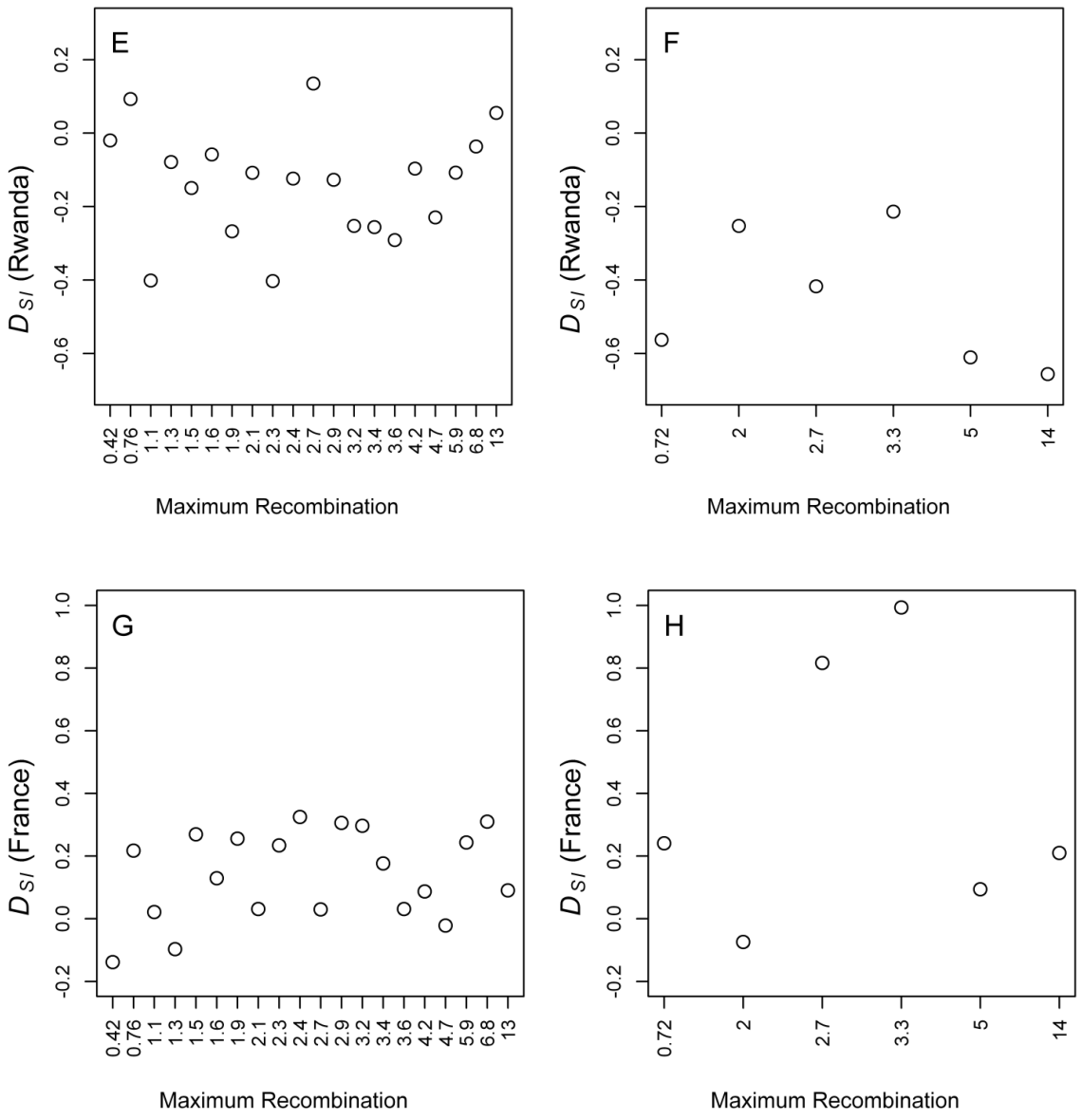

E
F
G
H
π_{SI} Rwanda
π_{SI} France
Maximum Recombination
0.000
0.005
0.010
0.015
0.020
0.025
0.42
0.76
1.1
1.3
1.5
1.6
1.9
2.1
2.3
2.4
2.7
2.9
3.2
3.4
3.6
4.2
4.7
5.9
6.8
13
0.72
2
2.7
3.3
5
14

Figure S14. Tajima's *D* at 4-fold degenerate sites and positions 8-30bp of introns ≤ 65bp in length (SI sites), binned by recombination rate. (A) *D* at autosomal 4-fold degenerate sites in regions where crossing over occurs (CDS-C) and in regions where crossing over does not occur (CDS-NC), Rwandan sample; (B) *D* at X-linked 4-fold CDS-C and CDS-NC, Rwandan sample; (C) *D* at autosomal 4-fold CDS-C and CDS-NC, French sample; (D) *D* at X-linked 4-fold CDS-C and CDS-NC, French sample; (E) *D* at autosomal SI sites, Rwandan sample; (F) *D* at X-linked SI sites, Rwandan sample; (G) *D* at autosomal SI sites, French sample; (H) *D* at X-linked SI sites, French sample.




E
F
G
H
D_{SI} (Rwanda)
D_{SI} (France)
Maximum Recombination
0.42
0.76
1.1
1.3
1.5
1.6
1.9
2.1
2.3
2.4
2.7
2.9
3.2
3.4
3.6
4.2
4.7
5.9
6.8
13
0.72
2
2.7
3.3
5
14

Chapter 3. Balancing selection maintains genetic polymorphism at a locus responsible for colour polymorphism in Gouldian finches, *Erythrura gouldiae*

Contributing authors: Benjamin C. Jackson, Kang-Wook Kim, Terry A Burke and Kai Zeng

Abstract

We investigate the population genetics of a locus underlying a striking, naturally occurring colour polymorphism in the Gouldian Finch, *Erythrura gouldiae*. Coalescent simulations and a modified implementation of the HKA test show that patterns of nucleotide diversity at the locus of interest are incompatible with the neutral model of molecular evolution, which suggests a role for balancing selection in maintaining colour polymorphism in these birds. In order to reach this conclusion we account for non-random sampling at our focal locus when trying to i) infer key population genetic parameters (e.g., the recombination rate), and ii) detect signatures of selection. Our results demonstrate the importance of correctly accounting for sampling bias before making conclusions about the evolutionary history of functionally important parts of the genome. We also find evidence for the action of GC-biased gene conversion in Z-linked intronic sequences in Gouldian finches.

Introduction

Associating genotypes with biologically relevant phenotypes offers a powerful framework for understanding the genetic basis of adaptation, and the maintenance of polymorphism in natural populations (Feder and Mitchell-Olds 2003; Dean and Thornton 2007; Mitchell-Olds et al. 2007). Such investigations might proceed in two stages. Firstly, by the identification of a region of the genome responsible for a phenotype, and secondly, by investigating patterns of genetic variation at those trait loci, using population genetic methods. The study of the *fast* and *slow* alleles at the alcohol dehydrogenase (*Adh*) locus in *Drosophila melanogaster* (Kreitman and Hudson 1991) offers the prototypical example of the use of population genetic data to test for evidence of natural selection, and led directly to the formulation of two widely-used statistical tests for the presence of non-neutral evolution, the HKA test (Hudson et al. 1987), and the McDonald-Kreitman test (McDonald and Kreitman 1991), with the conclusion that the *Adh* gene is under balancing selection (Kreitman and Hudson 1991).

More recently, population genetic data has been used to demonstrate convergent positive selection in European and African human populations at the *LCT* gene, which encodes lactase, by showing extended haplotype homozygosity around SNPs associated with lactase persistence (Tishkoff et al. 2007). Patterns of linkage disequilibrium and nucleotide diversity associated with an allele at another human gene, G6PD, show that it has been subject to positive selection because that allele offers protection from severe malaria (Saunders et al. 2005). Loci implicated in wing patterning in *Heliconius* show elevated levels of population differentiation between *Heliconius melpomene* races, which suggests these regions are under divergent selection (Nadeau et al. 2012). In *Arabidopsis*, comparison of polymorphism at 27 disease resistance genes with hundreds of other regions throughout the genome found evidence of widespread balancing selection at those resistance genes (Bakker et al. 2006).

However, some care must be taken when using population genetics methods to infer the action of selection. For instance, some studies have conditioned their sampling scheme on a focal locus (e.g. Saunders et al. 2005). This might be a locus identified as associated with a trait of interest, using quantitative trait loci mapping or genome-wide association studies, for example. Firstly, F_{ST}, which is defined as the proportion of genetic variation explained in allele frequency differences between populations (Wright 1951; Charlesworth 1998), must be 1 between two alleles at the focal locus (upon which sampling was conditioned) if the two ‘populations’ are defined by the alleles they carry

at that site (Figure 1A). This also implies that there will be a local elevation in differentiation between the populations at sites linked to the focal site. However, as this result is an artefact of the sampling scheme, and is expected even under neutral evolution, it cannot tell us anything about the strength of selection at this site (Figure 1A). Secondly, if the frequencies of sampled alleles at a locus do not represent the allele frequencies in nature, then this strategy presents the danger of obtaining misleading results using other statistics. Figure 1C; D demonstrates this problem for the case when two alleles are represented in a sample at 50%. In the example in Figure 1, in nature their frequencies are 33% and 67%. This disproportionate sampling imposes artificial genealogical structure at the focal site, which has implications for the significance of statistics like total diversity or Tajima's D (Tajima 1989a) if they are tested against a null model which does not take the sampling scheme into account. Tajima himself (Tajima 1989a) cautioned that "*The DNA sequences applied to this method must be a random sample from a population*". Pannell (2003) has shown that the value of Tajima's D is elevated in the presence of population structure. This pattern might be confused with the expectation under balancing selection (Charlesworth 2006).

Hudson et al. (1994) dealt explicitly with non-random sampling at a focal locus by subsampling in order to reproduce the allele frequencies observed in nature, before making inferences about the presence of natural selection(Hudson et al. 1994). However, this strategy may result in a reduction of power, especially if the allele frequencies in nature stray far from those in the sampling scheme (in which case more samples will be required to be discarded). In work on the G6PD gene in humans, Saunders and colleagues (Saunders et al. 2005) used previous allele frequency surveys to construct random samples with the appropriate allele frequencies from their own (non-random) data in order to estimate polymorphism and site frequency spectrum statistics. In addition, they used coalescent simulations conditioning on their estimates of the true derived allele frequency to get significance levels for observed levels of LD. They concluded from these analyses that the A- allele at G6PD has recently increased in frequency due to strong selection (Saunders et al. 2005). These examples highlight the importance of taking into account sampling schemes before trying to make inferences about the presence of natural selection using population genetic data.

The Gouldian finch, *Erythruria gouldiae*, is a North Australian endemic passerine bird, which exhibits a striking cheek patch colour polymorphism in nature (Southern 1945). Black-, red- and yellow-cheeked birds co-occur in natural populations at stable

frequencies (black: 70%, red: 30%, yellow: < 1%) (Franklin and Dostine 2000; in Pryke and Griffith 2009b). Cheek-patch colour is genetically controlled by two loci: one autosomal locus involved in carotenoid production, and one sex-linked locus involved in melanin production. The dominant allele at the autosomal carotenoid locus produces red-cheeked finches, whilst the recessive allele produces yellow-cheeked finches, provided there is at least one dominant allele at the sex-linked melanin locus. The recessive allele at the melanin locus produces black-cheeked finches (Brush and Seifreid 1968) (Figure 2). There is evidence from captive-bred birds that cheek patch colour is associated with a suite of behavioural and genetic traits, including behavioural dominance (Pryke and Griffith 2006; Pryke 2007), mate choice (Pryke and Griffith 2007; Pryke and Griffith 2009a) and post-zygotic isolating mechanisms (Pryke and Griffith 2009b), which suggests that the locus underlying these traits might be subject to some form of natural selection.

Kim et al (2016) have recently determined the genomic location of the sex-linked melanin locus (the *Red* locus) that defines black versus red/yellow phenotypes. By using linkage mapping with 35 Z-linked microsatellite markers and a pedigree consisting of 618 captive birds, and subsequently restriction-site associated DNA sequencing of 20 wild individuals, Kim et al (2016) mapped the *Red* locus to a position equivalent to the 46.5Mbp region of the Zebra finch Z chromosome. They further confirmed that two haplotypes around one diagnostic restriction site associated DNA (RAD) marker were associated with cheek-patch colour in Gouldian finches using allele-specific PCR genotyping of 161 wild individuals.

The *Red* locus lies in an intergenic region, between the genes *MOCS2* and *FST*. *FST*, which encodes the protein follistatin, was identified as a candidate gene underlying the phenotypic difference between colour morphs. Follistatin regulates hair and skin development in mammals (e.g. McDowall et al. 2008). It is antagonistic with TGF-β superfamily genes, which play a role in melanocyte differentiation, and also interacts with the MC1R signal pathway, which is involved in melanin production (Kim et al in prep, and references therein). However, no between-morph sequence differences were found within *FST*, which suggests that variation in a regulatory region controlling *FST* expression – putatively the *Red* locus – might underlie the phenotypic differences between morphs. Interestingly, the *Red* locus includes a putative transposable element insertion, which is not present in its nearest relative for which we have sequence data, the Zebra finch (Kim et al., in prep).

To further investigate the *Red* locus, 12 wild-caught female birds from a single population, six each of the red and black colour morphs, were sequenced for a contiguous ~100kbp stretch of the Z chromosome, spanning the *Red* locus. Because these birds are hemizygous at the *Red* locus, they represent six each of the recessive black (*b*) and dominant red (*R*) alleles. A test between red and black birds from this population, using 9 autosomal microsatellite markers, returned no evidence for any population structure (Kim et al, in prep).

In this work, we aim to investigate whether or not the patterns of genetic diversity observed within and between cheek patch colour morphs at the *Red* locus are the result of natural selection. To this end, we employ simulations to generate patterns of polymorphism under the neutral expectation, conditioning on our sampling scheme and the allele frequencies in the wild population at the *Red* locus. This is because standard population genetics approaches generally assume a random sample from a population(s), which is not the case in our dataset. To do so, we estimate population genetic parameters including the population mutation rate (θ) and the population recombination rate (ρ), taking into account demography and other non-selective forces, which we infer from Z-linked intronic loci unlinked to the *Red* locus. It is important to consider demography when testing for selection, because the effect of demographic events on genetic polymorphism may be mistaken for the effect of selection (Tajima 1989a; Tajima 1989b; Zeng et al. 2006). Similarly, GC-biased gene conversion (gBGC) has the potential to distort the site frequency spectrum (SFS) in such away which may lead to incorrect inferences of selection and/or demography (Eyre-Walker 1999; Duret and Galtier 2009; Zeng 2012). gBGC has been suggested to play a part in the evolution of bird genomes (Backström et al. 2013; Weber et al. 2014; Bolívar et al. 2015), so we attempt to quantify its effect in this study at the same time as making demographic inferences. We also implement a modified version of the HKA test (Hudson et al. 1987) which accounts for non-random sampling and recombination at one locus, as well as uneven lengths between loci.

We find evidence for non-neutral evolution at the *Red* locus from both our simulation studies, and the modified HKA test, and conclude that the observed patterns of polymorphism are most compatible with the action of balancing selection. Our theoretical consideration of the problems introduced by non-random sampling would be useful to future researchers designing sampling schemes to investigate loci underlying phenotypes of interest.

Materials and Methods

Genetic regions under consideration

We analysed a 99,669bp stretch of sequence from the *Red* locus in 12 female birds, 6 each from the black and red phenotypes. One additional sequence stretch (henceforth the 'flanking' region), approximately 7kbp long, was obtained for the same samples adjacent to the *Red* locus. The missing region in between the *Red* locus and the flanking region consists of approximately 5-10kbp.

We also chose 24 Z-linked loci elsewhere on the Z chromosome, far away enough from the *Red* locus that their diversity patterns should not be affected by non-random sampling with respect to the *Red* locus (Kim et al, *in prep*). These loci were assayed to obtain background information on patterns of polymorphism and differentiation on the Z chromosome, in order to be used as a reference. They consisted of single introns, 936-1135bp in length, and were sampled for between 11 and 12 female individuals, 10 (5 each of Red and Black) of which are also represented in our sample of the *Red* locus.

Estimates of DNA polymorphism and divergence

In order to compare the reference loci with the *Red* locus, we calculated π and Tajima's *D* on the full samples at these loci (*b* and *R* alleles combined) using custom scripts in R (R Core Team, 2014). We then calculated the same statistics within each allelic class (we define two allelic classes: one for each of the birds having the *b* or the *R* allele at the *Red* locus) at the reference loci and the *Red* locus, in order to compare within-allelic class polymorphism with the level of polymorphism for the combined sample. This was to ascertain, for instance, whether polymorphism at the *Red* locus might have been maintained by balancing selection (Charlesworth 2006). In addition, we calculated F_{ST} and D_{XY} between the two allelic classes at the reference and the *Red* loci using custom scripts in R. We used Weir and Cockerham's (Weir and Cockerham 1984) definition of F_{ST}, and the estimator provided by Hudson et al (Hudson et al. 1992). To combine information from multiple SNPs, we used equation (6) of Jackson et al. (2015) [see also (Bhatia et al. 2013)].

Below, we define statistics that refer to subsets of the data using subscripts: for example π_{red} refers to nucleotide diversity within the red allelic class, and π_{total} refers to nucleotide diversity with both the black and the red allelic classes combined.

When conducting point estimates of summary statistics *per locus* we removed sites with missing data in any individual. For sliding window analyses, when calculating F_{ST}, D_{XY} and π, we retained sites with missing data to maximise the number of SNPs in each window. However, because Tajima's *D* cannot be calculated with the existence of missing data, we removed individuals with missing data on a *per window* basis for this statistic. For all sliding window analyses, we used a window size of 1000bp with a step of 500bp.

Testing for a change in population size and gBGC

We used the matrix method of Zeng and Charlesworth (2009; see also Evans et al. 2014) to estimate the likelihood of a one-step change in population size, based on our 24 Z-linked reference loci. Briefly, this method uses the site-frequency spectrum to infer parameters of a two-allele model with reversible mutation, selection and/or gBGC, and changes in population size. We defined A/T and G/C as our two alleles. We define u as the rate at which A/T alleles mutate to G/C alleles, and v as the mutation rate in the opposite direction, and $\kappa = v/u$ as the mutation bias parameter. The selection coefficient, γ, is defined as $\gamma = 4N_e s$ where N_e is the effective population size and s is the selection coefficient against heterozygous carriers of the G/C allele in our case. To model a change in population size, we assume that the population in the past is at equilibrium with population size N_1, which then changes instantly to N_0 (this can be either an increase or a reduction in size) and remains in this state for t generations until a sample is taken from the population in the present day (Zeng and Charlesworth 2009; Haddrill et al. 2011; Evans et al. 2014). For each model, in order to ensure that the true MLE was found, we ran the search algorithm multiple times (typically 1000), each initialised from a random starting point. All the results reported below were found by multiple searches with different starting conditions. Chi-squared tests were used to evaluate statistical support for different models.

Testing for selection

We employ two widely-used methods for testing for the presence of natural selection. Firstly, we use coalescent simulations to generate polymorphism data under neutrality at a locus that is equivalent to the real *Red* locus in Gouldian finches. By assessing

whether the observed polymorphism at the *Red* locus is compatible with these simulations, we can accept or reject the null hypothesis of neutrality. We also used the HKA test (Hudson et al. 1987) to test for selection. More detailed methodology is presented in the sections below.

Considering non-random sampling at the Red *locus*

As mentioned in the Introduction, it is important to take into account non-random sampling with respect to the *Red* locus when testing for departures from neutrality using polymorphism data. This is because, even under neutrality, it is possible to obtain maximal F_{ST} values, as well as elevated π at a focal site if the sampling scheme is conditioned on that site, as in our case (see Figure 1). Thus, sampling scheme must be taken into account when carrying out, for instance, simulations to generate polymorphism data under the null hypothesis of neutrality. It is worth re-iterating that our reference loci should be unaffected by the non-random sampling at the *Red* locus, because they are unlinked to it (Kim et al, in prep). In order to replicate an observed sampling scheme for the purpose of statistical testing, we need to know the allele frequencies in nature. We used 161 birds, genotyped at the *Red* locus by Kim et al (in prep) for this purpose (Supplementary Table S1). Importantly, we do not know which allele (*b* or *R*) is ancestral, so it is necessary to consider both scenarios in turn; i.e. *b* is derived (derived allele frequency (DAF) = 0.856); or *R* is derived (DAF = 0.144) (see below for details about allele frequency estimation). This is in part because we cannot be sure which site(s) within the *Red* locus is causal (i.e. the 'focal' site represented in Figure 1A), and so cannot polarise that site by an outgroup sequence. In addition, we lack outgroup sequence data for a portion of the *Red* locus which coincides with a putative TE insertion in the Gouldian finch lineage, but which does not exist at the same location in the Zebra finch genome. As such, for the majority of the simulations and analyses below we arbitrarily set the focal site to the centre of the *Red* locus. However, our results are robust to different choices of this parameter (see Discussion).

We used the programs SelSim (Spencer and Coop 2004) and mbs (Teshima and Innan 2009) to simulate polymorphism linked to a focal site, at which a derived allele has arisen and spread by genetic drift, under neutrality, to the frequencies we observed at the *Red* locus (0.144 or 0.856). These programs allowed us to replicate the allele frequencies, sampling scheme and population genetic parameters observed at the *Red*

locus and obtain expectations under neutrality. In addition, mbs can model demographic changes.

Maximum likelihood estimates of allele frequencies

Testing for departures from neutrality at the *Red* locus requires information about both allele frequencies of the *b* and *R* alleles at the *Red* locus, denoted as p_b and p_R, respectively. A chi-squared test using the 161 wild birds genotyped at the *Red* locus did not detect evidence of departure from expected Hardy-Weinberg proportions (P = 0.816, df = 2) (Supplementary Table S1). We then calculated the ln-likelihood of p_b and p_R as

$$L(D|p_b, p_R) = 2X_{bb}\ln(p_b) + X_{bR}\ln[2p_b p_R] + 2X_{RR}\ln(p_R) + Y_b\ln(p_b) + Y_R\ln(p_R) \quad (1)$$

Where D represents the genotypes of the 161 birds, X_{bb} is the observed count of males homozygous for *b*, X_{bR} is the observed count of heterozygotes in males, X_{RR} is the observed count of males homozygous for *R*, Y_b is the observed count of females hemizygous for *b* and Y_R is the observed count of females hemizygous for *R*. In agreement with previous reports that yellow-cheeked birds are extremely rare, none were captured in Kim et al.'s experiment. Subject to maximising eq. (1), the constraints $p_b + p_R = 1$ and $0 \leq p_b \leq 1$, our maximum likelihood estimates of the allele frequencies for the *b* and *R* alleles were 0.856 (95% CIs: 0.808, 0.897) and 0.144 (95% CIs: 0.103, 0.192) respectively.

Estimating θ under a change in population size

We sought to obtain an estimate of θ, which is defined as $4N_e u$, where N_e is the effective population size, and u is the mutation rate per site, under an inferred change in population size obtained from the method of Zeng and Charlesworth (2009; see above) (see Results). To do so, we wanted to determine a value of θ such that the expected level of diversity under this population-scaled mutation rate, given the inferred one-step

change in population size, equals the mean value of π_{total} observed at our 24 reference loci (Table 1).

The expected value of diversity at a neutral site is $\theta E(X)$, where X is the time in units of $2N_0$ generations, and N_0 is the population size after the change, to the most recent common ancestor for two randomly chosen alleles taken from the current population. If an expression of $\theta E(X)$ can be found, we can estimate θ using

$$\hat{\theta} = \frac{\pi_{obs}}{E(X)} \tag{2}$$

The coalescent rate in $[0, \tau]$ is 1 and that in $[\tau, \infty]$ is $a = N_0/N_1$, where τ is the time in units of $2N_0$ generations between the time of the inferred change in population size and the present (when a sample was taken from the population), N_0 is the population size in the present and N_1 is the population size before the inferred change in population size. Furthermore, the probability of coalescing in $[\tau,\infty]$, referred to as c, can be calculated as

$$c = \int_{\tau}^{\infty} e^{-t} dt = e^{-\tau} \tag{3}$$

Thus, the probability density function, $f(t)$, for the coalescent time X is

$$f(t) = \begin{cases} e^{-t}, & \text{if } 0 \le t \le \tau; \\ cae^{-a(t-\tau)}, & \text{if } t > \tau. \end{cases} \tag{4}$$

Thus, $E(X)$ can be calculated as

$$E(X) = \int_0^{\tau} te^{-t}dt + \int_{\tau}^{\infty} tcae^{-a(t-\tau)}dt \quad (5)$$
$$= 1 - e^{-\tau}(1+\tau) + \frac{e^{-\tau}(1+a\tau)}{a}$$

Replacing τ and a in eq. (5) with the MLEs from the Zeng and Charlesworth (2009) method, we obtain $E(X)$, which is then inserted into Eq (2) to obtain $\hat{\theta}$.

Estimating ρ, the population recombination rate

We inferred the population recombination rates, ρ ($4N_e r$, where N_e is the effective population size, and r is the recombination rate per site), using polymorphism. For the reference loci, LDhat (Auton and McVean 2007) or LDhelmet (Chan et al. 2012) were employed. Because our sample at the *Red* locus is not a random representation of the population, we refrained from obtaining estimates of ρ at the *Red* locus using these methods. To get around this difficulty, we estimated the R_M statistic of Hudson and Kaplan (1985), which tests for the number of recombination events that can be parsimoniously inferred from a sample of DNA sequences, using the computer package *RecMin* (Myers and Griffiths 2003) (The R_h statistic of Myers and Griffiths gave essentially the same results; data not shown). Importantly, provided that there has been no recurrent mutation, R_M is non-zero only when there have been recombination events in the history of the sample (Myers and Griffiths 2003; Supplementary Table S2). Thus, R_M should be a robust test for the presence of recombination.

The following procedure was used to obtain ρ values that were likely to be compatible with data at the locus. We first simulated a neutral locus equivalent to the *Red* locus using mbs (Teshima and Innan 2009), conditioning on our sampling scheme (six each of *b* and *R* alleles, from a wild allele frequency distribution of 0.856 and 0.144, with each allele being treated as derived in turn). To generate neutral variants, the simulations were conducted using the estimate of θ from our 24 reference loci under the size change model (see the previous section), and incorporating the inferred demographic expansion (see Results). A range of values of ρ, from 0 to 0.005bp^{-1}, were considered. We performed 1000 simulations for each combination of parameters. We then estimated the R_M on each simulated dataset and obtained the mean across replicates. The average values of R_M were plotted against ρ, and ρ values with average R_M comparable to the observed R_M were regarded as compatible with the data. Note that ρ for the *Red* locus is

defined as $4N_0r$, where N_0 is the present day population size (see Results). However, the value of ρ was also comparable (same order of magnitude) when a constant-size population was assumed (Supplementary Figures S1; S2).

It should be noted that we primarily used R_M as a test of evidence of recombination at the Red locus, and that the above method for estimating ρ is likely to be downwardly biased, as R_M is expected to miss many recombination events (Hudson and Kaplan 1985; Myers and Griffiths 2003). Nonetheless, this analysis provided valuable guidance on how to choose ρ in the presence of the non-random sampling. In addition, the value we obtained is comparable to those obtained by applying LDhat and LDhelmet to the reference loci and the flanking region (see Results). Finally, our results are robust to different choices of ρ (and to the case where we assume a constant-size population model). For instance, our results remained unchanged when using a value of ρ ten times smaller than that inferred from the above procedure (data not shown).

Coalescent simulations of the Red locus – under a population size change

In order to assess whether the patterns of genetic polymorphism at the *Red* locus are compatible with purely neutral processes we conducted further simulations using mbs (Teshima and Innan 2009). We simulated a 99,669bp region (equivalent in length to the *Red* locus), using the estimate of θ from our 24 reference loci under the size change model (Table 2), and the estimate of ρ from the method described above (see Results).

An advantage of using the maximum-likelihood methods described above is that it is possible to incorporate uncertainty surrounding the derived allele frequency (DAF) used to obtain the probabilities of tests of neutrality. We drew values of DAF from the posterior distribution of allele frequencies, which is proportional to equation (1). To do this, we used a rejection-sampling algorithm to sample values of p from its posterior distribution. We replicated the missing data in our observed dataset by masking the same sites in our simulated datasets before calculating summary statistics. We carried out two sets of simulations, treating each allele (*b* and *R*) as derived in turn.

A modified HKA test incorporating non-random sampling and recombination

We sought to modify the HKA test (Hudson et al. 1987) for four reasons. Firstly, our sample of the *Red* locus is non-random, which is not allowed for in the original

formulation of the test. Secondly, *p*-values from the X^2 distribution are not robust when we have unequal locus sizes (see Supplementary Material; Supplementary Figure S3). Thirdly, simulation-based *p*-values assuming no recombination in the *Red* locus may be too conservative, given its size and clear evidence of intra-locus recombination. Fourthly, a sliding window approach may help to identify candidate sites by contrasting the results of the test for different regions of the *Red* locus whilst retaining the same reference loci as a comparison.

We also carried out simulations to assess the effect of a one-step change in population size on the HKA test, and found that, in the presence of a demographic expansion equivalent to that which we inferred from our data, not directly accounting for the size change renders the HKA test conservative (Supplementary Figure S4). Therefore we chose not to incorporate the change in population size in our HKA analysis.

We first set out to estimate the expectation and the variance of the number of segregating sites within a window of the *Red* locus, given the sampling scheme and the location of the focal site. Assuming the infinite-sites model and defining θ *per window* as $4N_e u$, the mean and variance of the number of segregating sites in this window, *S*, are given by

$$E(S) = \theta E(T) \tag{6}$$

and

$$Var(S) = E[Var(S|T)] + Var[E(S|T)] = \theta E(T) + \theta^2 Var(T) \tag{7}$$

where *T* is the sum of the total branch length across all nucleotide sites in the window. The above equations have made use of the fact that, given *T*, *S* is Poisson distributed with mean θT.

In order to obtain $E(T)$ and $Var(T)$ for our sampling scheme and the recombination rate at (a window of) the *Red* locus, we used the following algorithm:

1) Set θ_{sim} *per window* to 1.0 (the choice of θ_{sim} is unimportant)
2) For *i* from 1 to 1,000,000

   Do

   a) Use SelSim and θ_{sim} to generate a random sample conforming to the *Red* locus (in terms of derived allele frequency, sample size at each

allele class, recombination rate, length and the site at which sampling is conditioned upon in the centre of the locus)

b) Record the number of segregating sites in each window of the *Red* locus and store it in X_i

Done

To conduct SelSim simulations of the *Red* locus we simulated a 99,669bp region, placing the focal site in its centre (although this choice does not materially affect the results), and using the same sampling scheme as in the observed data (six each of the two allele classes). We set $\rho = 0.001$ (see Supplementary Material – estimating ρ under equilibrium), the value of θ (*per window*) as 1 and used point estimates of the two DAFs (0.144 and 0.856). As with the position of the focal site, our results were robust to this value of ρ. We did not sample DAFs from their posterior distributions primarily because it was computationally intractable to perform the large number of simulations needed to estimate $E(T)$ and Var(T). Furthermore, the estimates of DAF have fairly small 95% confidence intervals.

Using the number of segregating sites (X_is) obtained from the simulations described above, we obtained estimates of $E(T)$ and $Var(T)$, denoted as μ_T and σ_T^2 respectively, as follows

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{8}$$

where n is the number of simulations performed in total

$$\mu_T = \bar{X}/\theta_{sim} \tag{9}$$

$$Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{10}$$

$$\sigma_T^2 = \frac{Y^2 - \theta_{sim}\mu_T}{\theta_{sim}^2} \tag{11}$$

Since μ_T and σ_T^2 were treated as known quantities in the analysis below, to minimise statistical noise, we estimated them using 10^6 simulations.

Now Equations (5) of Hudson et al (1987) for the case when we have polymorphism data from only one species can be rewritten as

$$\sum_{i=1}^{M-1} s_i + s_R = \sum_{i=1}^{M-1} C(n_i)\hat{\theta}_i + \hat{\theta}_R\mu_T,$$

$$\sum_{i=1}^{M} D_i = \left(\hat{Y} + 1\right)\sum_{i=1}^{M} \hat{\theta}_i, \tag{12}$$

$$s_i + D_i = \hat{\theta}_i\{\hat{Y} + 1 + C(n_i)\}, \qquad i = 1, \ldots, M-1.$$

where M is the total number of loci, among which the first M - 1 are our reference loci, and the M^{th} locus is the focal window of the *Red* locus, $C(n_i)$ is $\sum_{j=1}^{n_i-1} 1/j$ where n is the sample size at the i^{th} locus, D_i is the observed divergence between the ingroup data and the outgroup data, and Y is the divergence time between the ingroup and outgroup in units of $2N$ generations. In Eq. (12), hats have been added to parameters that are to be estimated, whereas all other symbols represent values obtained from the sequence data.

We carried out the test above as a sliding window across the *Red* locus, using a window size of 1000bp and a step size of 500bp. For each given window, we solved the M + 1 unknowns in Eq. (12) using a custom script in R, and calculated the X^2 statistic as in (Hudson et al. 1987), except that we obtained the mean and variance of tree lengths (T) from equations (9) and (11) above, then used our estimate of θ along with equations (6) and (7) to obtain the mean and variance of S. Details of how to solve Eq (12) for the classic HKA test, as well as example R code, are provided in the Appendix to this chapter.

In order to test the significance of X^2 we used two methods, i) the chi-squared distribution with $M - 1$ degrees of freedom and ii) by conducting coalescent simulations.

The procedure for conducting coalescent simulations for the reference loci is the similar to that described in Hudson et al. (Hudson et al. 1987). The main difference is that, for a window of the *Red* locus, we simulated polymorphism data using the algorithm given above, but using our estimate of θ in place of θ_{sim}. We simulated divergence at each locus (D_i) as

$$D_i = Pois\left(\hat{\theta}_i \hat{Y}\right) + Geom\left(\frac{1}{1+\hat{\theta}_i}\right) \quad (13)$$

Where $\hat{\theta}_i$ and $\hat{Y}$ were estimates obtained by solving Eq. (12) above (Hudson 1990). To estimate the p-value, we carried out 1,000 simulations for each window and estimated the probability that the value of X^2 from the simulation procedure is greater than or equal to our observed value of X^2.

We carried out the above procedure twice, once each for a DAF of 0.144 and of 0.856 at the focal site.

For the 2781bp portion of the *Red* locus for which we do not have divergence data, which coincides a transposable element (TE) insertion specific to the Gouldian finch lineage (and thus not present in the Zebra finch outgroup), we estimated divergence using 3kbp (1.5kbp from each side) of the *Red* locus immediately adjacent to the missing data. This procedure returned a value for divergence of 52.1kbp^{-1} which is greater than the mean value of the windows of the *Red* locus for which outgroup sequence data is available (46.1kbp^{-1}), making our HKA test conservative for this region in comparison to using a locus-wide average value of divergence.

Results

Patterns of DNA polymorphism and divergence at the Red *and reference loci*

Recall that we defined two allelic classes in our data, one for each of the two alleles at the *Red* locus. As all birds in our sample of the *Red* locus are female and therefore hemizygous at the *Red* locus, birds with the black phenotype will have *b* allele, and similarly, birds with the red phenotype will have the *R* allele at the *Red* locus. Mean values of summary statistics for the *Red* locus and the reference loci, which are unlinked to the *Red* locus are summarised in Table 1. The mean π at the reference loci is 0.0032, with no difference between allelic classes, or between the within-allelic class values and the value of π for the allelic classes combined. The point estimates of π at the *Red* locus within either the red (π_{red} = 0.0008) or the black (π_{black} = 0.0015) allelic classes are lower than π at the reference loci, but the value for both allelic classes combined at the Red locus (π_{total} = 0.0092) is markedly higher than π at the reference loci (π_{total} = 0.0032) (Mann-Whitney U test, $p < 0.001$). Tajima's *D* is moderately negative at the reference loci (-0.290, -0.809 and -0.601 for D_{red}, D_{black} and D_{total}, respectively) as well as within allelic classes at the *Red* locus (-0.199 and -0.911 for D_{red} and D_{black}, respectively), but is notably positive when both allelic classes are combined at the *Red* locus (D_{total} = 2.013). The latter value is significantly higher than 0 under the standard neutral model ($p_{coalescent}$ = 0.002), and significantly higher than D_{total} at the reference loci (Mann-Whitney U test, $p < 0.001$). F_{ST} at the *Red* locus (F_{ST} = 0.935) is both significantly greater than 0 ($p_{permutation}$ = 0.002), and significantly greater than F_{ST} at the reference loci (F_{ST} = 0.037) (Mann-Whitney U test, $p < 0.001$). D_{XY} is also significantly greater at the *Red* locus (D_{XY} = 0.0149) than at the reference loci (D_{XY} = 0.0033) (Mann-Whitney U test, $p < 0.001$). Polymorphism within allelic classes at the flanking locus is often intermediate between the values for the reference loci and the *Red* locus (Table 1).

We carried out a sliding window analysis of polymorphism and divergence at the *Red* locus (Figure 3). Some notable features are apparent from this analysis. Firstly, F_{ST} between the *b* and *R* alleles is nearly maximal (approaching 1) for a large stretch of sequence across the centre of the *Red* locus (Figure 3A). This region also exhibits elevated D_{XY} between *b* and *R* alleles (Figure 3B), Tajima's *D* (Figure 3C) and π_{total} (Figure 3D) compared to the periphery of the *Red* locus, the reference loci, and the flanking region (Table 1; Figure 3). Values of D_{XY} between *b* and *R* alleles at the *Red* locus also represent a significant proportion of the total divergence between Gouldian

Finch and Zebra Finch at the *Red* locus (Figure 3B). Taken at face value, these patterns, as well of those from the point estimates, are suggestive of the action of balancing selection. In what follows, we set out to explore to what extent they were caused by the non-random sampling scheme.

Evidence for a demographic expansion and gBGC

The one-step population size change model of Zeng and Charlesworth (2009) fitted the data significantly better than the equilibrium model, given the data from our reference loci ($X^2 = 38.76$, d.f. = 2, $p < 0.001$; Table 2). The time of the size change, τ was estimated as 9.63e-2 into the past, and the population was inferred to have undergone an expansion (forward in time) at this point by a factor of $a = 6.29$ (Table 2). Including a second size change did not result in a significant increase in the log-likelihood of the model over a single size change ($X^2 = 0.84$, d.f. = 2, $p = 0.66$; full data not shown). The one-step size change model was also a significant improvement on a one-step size change model with gamma set to 0 ($X^2 = 8.54$, d.f. = 1, $p = 0.003$; Table 2) i.e., a model incorporating selection (or some other force which favours one of our two allelic classes, A/T vs. G/C) fits the data better. The inferred value of γ was -0.564, which indicates a force favouring G/C alleles over A/T alleles. The mutation bias parameter, $\kappa = \frac{GC \rightarrow AT}{AT \rightarrow GC} = 3.2$ under the best fit model. This result is close the value reported in Flycatchers by Bolívar et al. (2015) of 3.41. The likelihoods and inferred parameters of these models are presented in Table 2.

Estimates of theta and recombination rate

Using the maximum likelihood estimates of τ (the time to the size change in units of $2N_0$ generations; see above) and a (the factor by which the population increased) we can obtain $E(X) = 0.2362$ from equation (5). Since $\pi_{obs} = 0.0032$ (Table 1), from equation (2) we have $\hat{\theta} = 0.0135$.

We estimated the minimum parsimonious number of recombination events (R_M) at the *Red* locus as 58. By simulating a region equivalent in length, sample size and allele frequencies, and using a range of values of ρ, we inferred that a value of $R_M = 58$ is compatible with $\rho \approx 0.004$, where $\rho = 4N_0r$, with the inferred demographic expansion (Supplementary Figure S1). We assumed a value of $\rho = 0.004$ for all analyses

incorporating the expansion below. This estimate lies within the estimates of ρ bp^{-1} at the control loci (Supplementary Table S3).

Testing the neutral hypothesis using coalescent simulations

As a first test, we examined whether patterns of diversity at the *Red* locus, as summarised by sliding-window plots of F_{ST}, D_{XY}, π and Tajima's *D* could be explained by a neutral model with a non-random sampling scheme and the inferred demographic expansion, as well as the inferred values of θ and ρ. Assuming a DAF of 0.144 at the focal site (the *R* allele is derived), the observed data at the *Red* locus lies above the 95% confidence intervals (CIs) of our coalescent simulations for many windows, for a range of summary statistics that incorporate both allelic classes (Figure 4). For the statistics calculated using within allelic classes (D_{red}, D_{black}, π_{red}, π_{black}), in most cases, the observed data generally fall within the 95% CIs of the simulated distributions (Supplementary Figure S5), probably in part caused by their higher variance because they were calculated on smaller sample sizes. The exception was π_{red} which lies below the simulated 95% CIs of π_{red} (Supplementary Figure S5D). Balancing selection is expected to lead to reduced diversity at neutral sites linked to the site under selection within the less frequent allelic class, but on a much smaller scale relative to the increase in polymorphism when allelic classes are combined (Nordborg 1997; Charlesworth 2006). These predictions do not account for demographic changes, however, such as the one we suggest has occurred in the Gouldian finch and consequently, it is unclear to us whether the observed level of π_{red} is compatible with the action of balancing selection alone. These results are qualitatively identical for a DAF of 0.856 (Supplementary Figure S6). Overall, these simulations suggest that there are multiple aspects of the data that cannot be explained by the sampling scheme alone.

Testing the neutral hypothesis using the modified HKA test

Analysis of the *Red* locus based on R_M gave clear evidence of recombination. By carrying out a sliding window implementation of the test, it is possible to contrast the results of the test for different regions of the *Red* locus whilst retaining the same reference loci as a comparison. In this way, we were able to identify regions showing the most unusual patterns of polymorphism.

Our sliding window implementation of the HKA test, taking into account non-random sampling and recombination at the *Red* locus, reveals significant departures from neutrality for multiple windows of the *Red* locus (Figure 5). *P*-values for both DAFs of 0.144 and 0.856, and both the chi-squared and the simulation methods are all in close accord. The most significantly departed region coincides with the region of the *Red* locus for which we do not have divergence data to Zebra finch due to the insertion of a putative TE in the Gouldian finch lineage, and at which the levels of polymorphism within the Gouldian finch lineage are highest (Figure 3B). However, we obtain similar results elsewhere in the *Red* locus if we treat the TE region as the single focal site (by removing polymorphism data from that region) (Supplementary Figure S7).

Discussion

Investigating polymorphism and divergence data in regions underlying phenotypic variation offers a powerful approach to understanding which evolutionary processes may have driven this variation. However, as we have seen above, care should be taken to ensure that sampling schemes are designed to avoid the potential biases associated with conditioning on known focal sites. Some studies have dealt with the problem of non-random sampling by constructing subsamples which represent natural allele frequencies (Hudson et al. 1994; Saunders et al. 2005). Given that there is fast-increasing power to detect genes underlying traits of interest, our consideration of the importance of sampling scheme here should be of interest to a wide range of researchers.

Defining populations in terms of the allele they carry at a focal site may reduce the statistical power of tests. This is particularly true for F_{ST} here, with values close to the maximum value of one in regions neighbouring the focal site, making it hard to distinguish whether the high observed F_{ST} between the Red and Black allelic classes is due to the way in which populations are defined, or is a result of long-term balancing selection (Figure 4). From the simulated data shown in Figures 4 and S6 it is also apparent that the diversity patterns simulated under neutrality closely linked to the focal site may be very similar to what we expect under balancing selection. Total diversity and Tajima's D are both elevated because of the bias introduced by sampling non-randomly with respect to the focal site.

In our case, we inherited a sample size that we judged too small to reduce further in order to represent natural allele frequencies, which would have meant discarding five out of six red genotypes. Consequently we took a different approach: we explicitly incorporated our sampling scheme into tests of the neutral model of molecular evolution and found evidence for non-neutral evolution at the Gouldian finch *Red* locus.

Genetic diversity at the Red locus is consistent with a long-term stable polymorphism

Patterns of genetic diversity at the *Red* locus are compatible with a long-term, stable polymorphism. F_{ST} between morphs is nearly maximal across much of the region and D_{total} is strongly positive. D_{XY} between morphs and π_{total} are much elevated compared to the reference loci, despite divergence to Zebra finch being comparable at the *Red* locus and our reference loci (Figure 3B; Supplementary Figure S8). In fact, mean D_{XY} between the *b* and *R* alleles at the *Red* locus is 0.0149 (Table 1), which is approximately a third of the mean value of the interspecific D_{XY} between Gouldian finch and Zebra finch at the *Red* locus (0.047), and is ~4.5x the value of π_{total} at the reference loci. At its highest point (in the putative TE insertion) D_{XY} between the *b* and *R* alleles reaches the mean value of interspecific divergence (Figure 3B). This suggests that these alleles are much older than an average polymorphism in the Gouldian finch Z chromosome, and implies that the *b* and *R* alleles may have been maintained at the *Red* locus for a substantial amount of time. All of these signals are over and above what is expected from the effect of non-random sampling alone.

Our results are robust to several key assumptions in the analyses we present above (i.e. the recombination rate, and the position of the focal site; see below). Additionally, they remain largely unchanged whether or not we consider demography. Carrying out coalescent simulations under the equilibrium situation (no population expansion) returned qualitatively the same results as the situation with a demographic expansion (Supplementary Figures S9, S10). In general, under the constant-size model, the simulated distributions of polymorphism summary statistics have larger variance than under the expansion model, such that the discordance between the observed data and the expectation under neutrality is greater under the expansion model. For instance, π_{red} does not lie outside the 95% CIs of our simulations under equilibrium, but does under the expansion model. This serves to highlight the importance of using realistic demographic models when implementing tests for natural selection.

Our results are also reasonably insensitive to the location of the focal site. The location of the true focal site, i.e. the candidate position upon which selection may have acted, was unknown to us, because there are hundreds of fixed differences between the *b* and *R* alleles at the *Red* locus, none of which is individually identifiable as causal. In addition, the *Red* locus itself is intergenic, which means a protein-coding change cannot be implicated in the polymorphism. As such, for the majority of our analyses (in our coalescent simulations, as well as to obtain expectations for, and to test the significance of, the HKA test) we arbitrarily set the position of the focal site at the centre of the *Red* locus. For the coalescent simulations, moving the focal site to a point under the peak of polymorphism at approximately the 60kbp point along the *Red* locus (see Figure 3B) did not result in any qualitative difference in the significance of our results (data not shown). We discuss the effect of moving the focal site on our modified HKA test below.

Some alternative explanations for the large block of elevated F_{ST} values across the centre of the *Red* locus (Figure 3A) presented themselves to us. These include the presence of a chromosomal inversion (or some other factor reducing recombination) between red and black morphs; a recent introgression or admixture event that has introduced one of the two alleles; that there is more than one site under selection, or some combination of these possibilities. Kim et al (in prep) tested for the presence of chromosomal inversion breakpoints by carrying out long range PCR across the *Red* locus, without finding any evidence of them. In addition, we have good evidence of non-zero recombination across the entire *Red* locus from i) a linkage map of the Gouldian Finch Z chromosome (Kim et al. 2016); ii) direct observation of recombinant genotypes at the *Red* locus (K-W Kim, pers. comm.); and iii) our R_M assay. Additionally, our estimate of the $R_M\text{bp}^{-1}$ rate remained unchanged when we only considered the portion of the *Red* locus with elevated F_{ST}. Further, our results remained qualitatively unchanged if we use a value of ρ an order of magnitude smaller than that which we infer at the *Red* locus (data not shown).

It seems reasonable to exclude the possibility of a recent admixture event given that there is no evidence of population structure between red and black birds, either from the Z-linked reference loci presented in this study, or 9 autosomal microsatellite loci (Kim et al, in prep). This leaves the possibility that there might be multiple selected sites across the block of F_{ST}, which our simulations do not take into account. However, as can been seen from Figure 4, even if the entire *Red* locus were non-recombining (in which case the distribution of polymorphism expected under neutrality would be

represented by an extension of the CIs shown in Figure 4 at the position of the focal site, across the entire *Red* locus), there would still be obvious differences between the observed and the expected data.

These observations, together with the fact that the *b* and *R* alleles are much older than an average variant segregating in the reference loci (as judged by comparing divergence data between *b* and *R*, and between Gouldian finch and Zebra finch [Figure 3B]) suggest that *b* and *R* have been maintained in the population for a long period of time. The elevated π_{total}, Tajima's *D* and D_{XY} at the *Red* locus are all consistent with the predictions of balancing selection (Nordborg et al. 1996; Charlesworth 2006).

The transposable element in the Red locus shows the strongest signature of selection

According to both our sliding window coalescent simulations and the HKA test, the part of the *Red* locus that shows the strongest evidence for selection is the TE insertion and its surrounding regions. This TE insertion does occur in the Zebra finch genome, but not at the same location on the Z chromosome as in Gouldian finches, which suggests its insertion into the *Red* locus has occurred since the split between the Gouldian finch and Zebra finch lineages. It was entirely absent from other passerine birds for which genomic data is available: Medium ground finch, Collared flycatcher, and Great tit, as well as from the Chicken, according to BLAST searches. In the Gouldian finch, the TE insertion itself at the *Red* locus displays an unusually high level of D_{XY} (between red and black alleles) (Figure 3B), reaching, or even exceeding, the value of the divergence between Gouldian finches and Zebra finches calculated overall at the *Red* locus (Figure 3B; Supplementary Figure S8), which suggests that it differentiated the *b* and *R* alleles for a long period of time, although the lack of divergence data makes it hard to take local variation of mutation rate into account.

Because of the lack of divergence data from the TE insertion region, we were interested in evaluating the robustness of our implementation of the HKA test by removing windows including the TE from the test, and moving the focal site to its position. The results of this test were very similar to the original test, for windows excluding those which incorporated the TE (Figure 5 and Supplementary Figure S7). However, only 6 and 3 windows are significant ($p < 0.05$) according to simulation from DAFs of 0.144 and 0.856, respectively, out of 192 windows in total – a total which we might expect to find merely by chance. This may be due to the fact that our test is rendered overly

conservative by assuming a constant population size (Supplementary Figure S4). These HKA analyses, with and without the TE, nevertheless serve the useful purpose of showing regions of the *Red* locus that depart most strongly from neutral expectations. As with the results from our coalescent simulations (Figure 4), the TE insertion was identified as the primary candidate (Figure 5).

We were also interested in extending the classic HKA test to use π as the measure of within-species polymorphism as opposed to the number of segregating sites (S), as in Kreitman and Hudson (1991). Although the variance associated with π is greater than that with S (which leads to less power to reject the null hypothesis of neutrality, all else being equal), the test using π is more sensitive to among-region differences in the case where polymorphism differs between regions, as in our case (Kreitman and Hudson 1991). The test using π was largely in agreement with the test using S, and returned moderately more significant results (Supplementary Figure S11). The methods associated with this test are presented in the Supplementary Material.

Evidence for GC-biased gene conversion

Using our reference loci, and the method of Zeng and Charlesworth (2009), we found evidence that G/C alleles are favoured over A/T alleles, despite an AT-bias in mutation, the magnitude of which was similar to another recent study in birds (Bolívar et al. 2015). Because the sequences we used to conduct these tests were intronic, this suggests the action of GC-biased gene conversion (gBGC) as opposed to selection for preferred codons, for instance. gBGC has been more thoroughly investigated in mammals than in birds, and its underlying mechanisms are best characterised in yeast (Duret and Galtier 2009), but its effect on avian genome evolution may be important (Bolívar et al. 2015). These data provide evidence in support of this.

Conclusions

By accounting for sampling scheme and using a set of reference loci to infer key population genetic parameters, we have demonstrated that patterns of genetic diversity at the *Red* locus, which is implicated in a colour polymorphism in Gouldian finches, are incompatible with neutral evolution. The observed patterns are, in the main, compatible with the action of balancing selection, although it remains unclear where the target of

selection is, and what differences in fitness there are between individuals with different genotypes. The theoretical considerations which were necessitated by our non-random sampling scheme should inform future workers on the importance of sampling scheme and the assumptions of commonly used population genetic tests. We also provide evidence for the action of GC-biased gene conversion on the Gouldian finch Z-chromosome.

Tables

Table 1. Point estimates of polymorphism statistics at the *Red* locus and flanking locus, and mean values at the reference loci. π – nucleotide diversity; D_{Taj} – Tajima's D; F_{ST} – differentiation between the Red and Black allelic classes; D_{XY} – mean pairwise diversity between the Red and Black allelic classes. Figures in brackets denote the standard error, where applicable.

Region	No. of loci	Mean locus length (bp)	Within Allelic Class			Classes Combined		Between Classes	
			Class	π	D_{Taj}	π	D_{Taj}	F_{ST}	D_{XY}
Reference	24	1041.5	Red	0.0032 (0.0005)	-0.290 (0.1231)	0.0032 (0.0004)	-0.809 (0.1243)	0.037 (0.0207)	0.0033 (0.0004)
			Black	0.0032 (0.0004)	-0.601 (0.1549)				
Red	1	99669	Red	0.0008	-0.199	0.0092	2.013	0.935	0.0149
			Black	0.0015	-0.911				
Flanking	1	6279	Red	0.0026	0.077	0.0025	-0.523	-0.091	0.0024
			Black	0.0027	0.234				

Table 2. Ln-likelihood results and parameter estimates from the method of Zeng and Charlesworth (2009) to test for a change in population size, and for non-neutral evolution. Parameters: u – the population-scaled AT→GC mutation rate; v – the population-scaled GC→AT mutation rate; γ – the population-scaled selection coefficient favouring AT alleles; a – the ratio of population sizes N_1/N_0 where N_0 is the population size before the change and N_1 is the population size after the change, forward in time; τ – the time in units of $2N_0$ generations into the past at which the inferred change in population size occurred.

Model	Max ln-likelihood	-2Δ lnL[a]	u	v	γ	a	τ
Equilibrium	-1.7526e+04		1.69e-03	6.14e-03	-6.93e-01	n/a	n/a
Size change 1 step, $\gamma = 0$	-1.7510e+04	38.75986	1.31e-03	2.41e-03	n/a	6.55e+00	9.97e-02
Size change 1 step	-1.7506e+04	8.543067	1.02e-03	3.26e-03	-5.64e-01	6.29e+00	9.63e-02

[a] Two times the difference in log-likelihoods between this model and the next worst model (the model above in the table).

Figures

Figure 1. Schematic representation of the effect of non-random sampling on population genetic summary statistics. A, Defining populations based on known variants at a causal locus results in an F_{ST} of 1 at the focal locus, and in elevated values of F_{ST} at tightly linked neutral variants. **B**, taking an equal sample size (blue lines represent the tree of sampled alleles) from two allelic classes when their allelic frequencies in nature are 33% and 67%, compared to the case when the sample is proportional to the natural allele frequencies, as is the case in panel **C**.

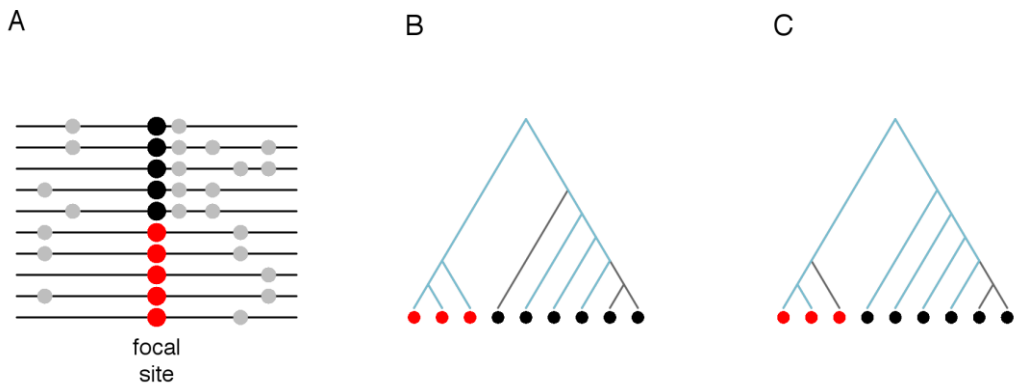


Figure 2. Cheek-patch colour phenotypes associated with genotypes at the Z-linked *Red* locus and the *Yellow* autosomal locus, which together control cheek patch colour. All possible genotypes at the *Yellow* locus (rows) and the *Red* locus (columns) result in either black, red or yellow phenotypes (cell colours). As the *Red* locus is sex-linked, females are hemizygous.

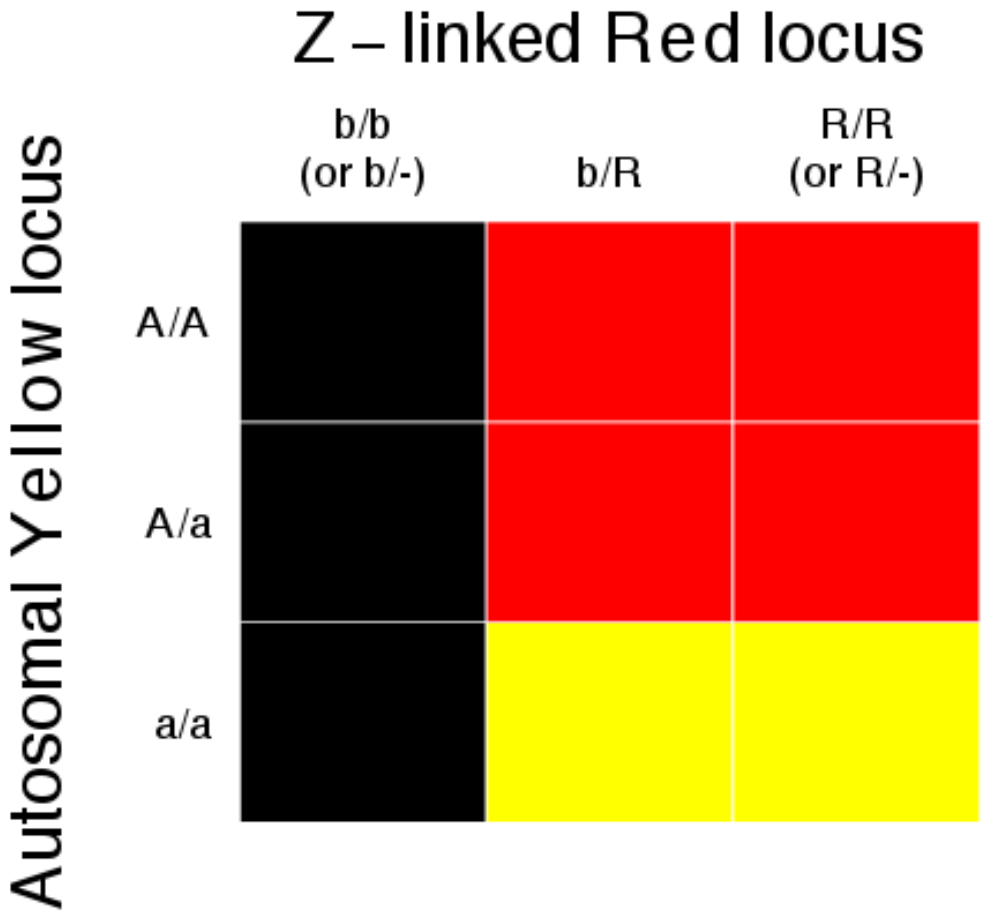


Figure 3. Sliding window evaluation of polymorphism and divergence at the *Red* locus, with comparison to 24 Z-linked introns that serve as neutral reference loci. **A**, F_{ST} between Black (*b*) and Red (*R*) allelic classes (black line) at the *Red* locus and the mean between *b* and *R* allelic classes at the reference loci (blue dashed line); **B**, D_{XY} between *b* and *R* allelic classes (black solid line) and between Gouldian Finch and Zebra Finch (orange solid line) at the *Red* locus, and the mean values of the same statistics at the reference loci (blue dashed lines); **C**, Tajima's D for *b* and *R* allelic classes combined; **D**, nucleotide diversity for *b* and *R* allelic classes combined (grey line), the *b* allelic class alone (black line), the *R* allelic class alone (red line) and the mean value for *b* and *R* allelic classes combined at reference loci (blue dashed line). The shaded grey bar indicates the region at which there is a putative TE insertion in the Gouldian Finch lineage.

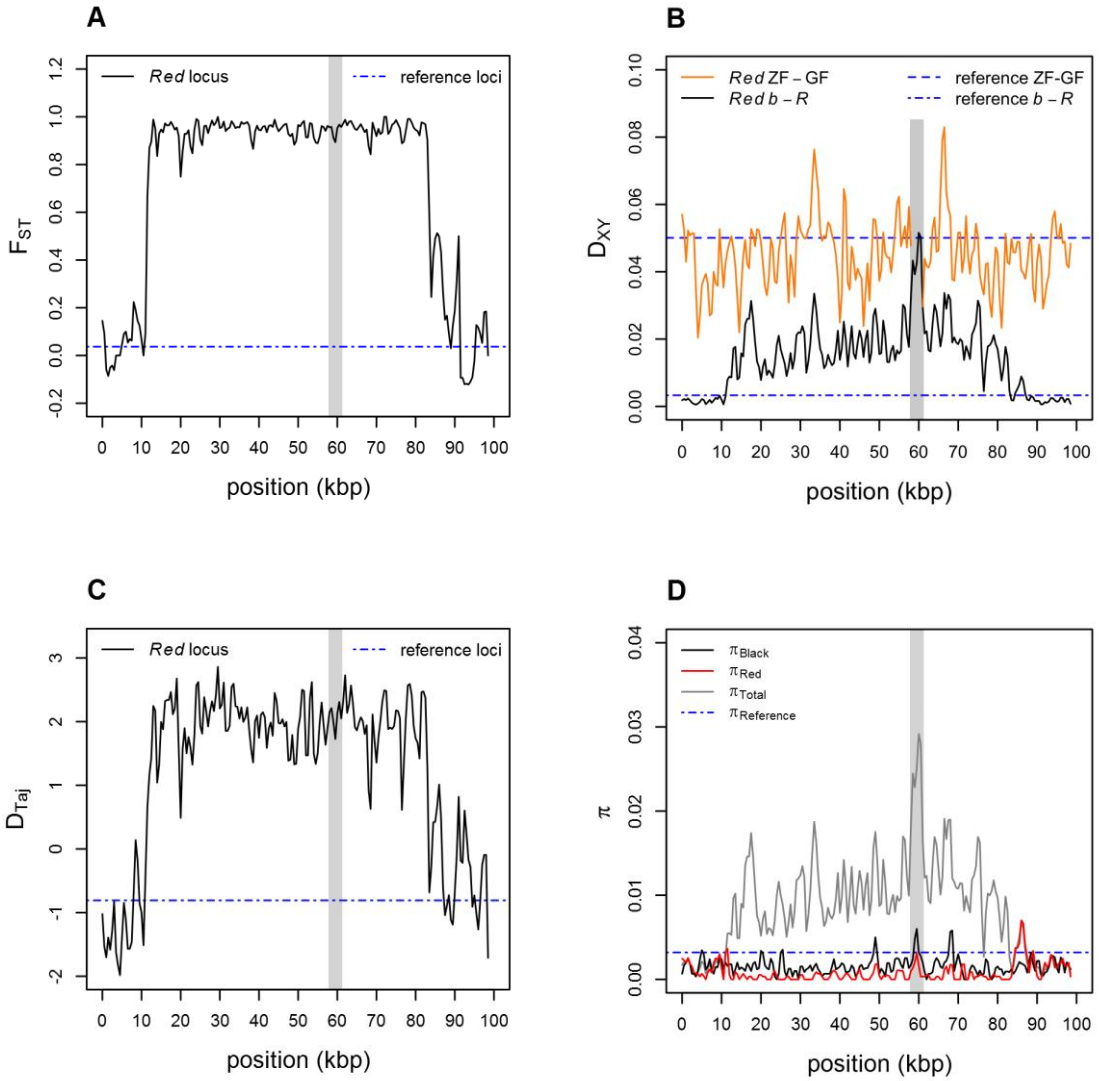


Figure 4. A sliding window comparison of observed polymorphism at the *Red* locus and polymorphism simulated under the standard neutral model of molecular evolution with a derived allele frequency of 0.144, taking into account the observed sampling scheme and recombination rate and incorporating a single change in population size. The focal site was placed at the centre of the simulated locus. Solid lines represent the observed data. The grey shaded region represents the space encompassed by the 95% confidence intervals derived from the simulations. **A**, F_{ST} between both allelic classes at the *Red* locus; **B**, D_{XY} between both allelic classes; **C**, Tajima's D for both allelic classes combined; **D**, nucleotide diversity for both allelic classes combined.

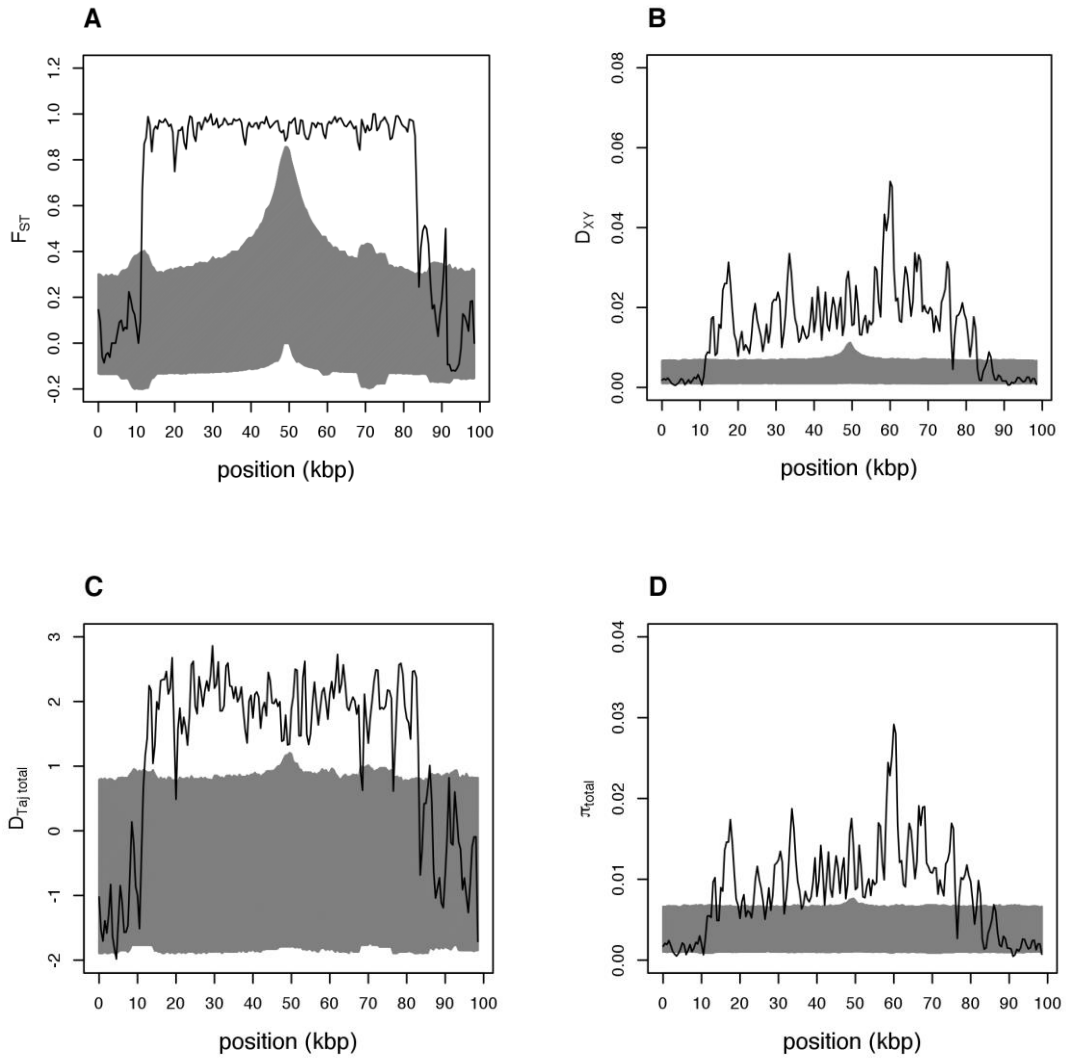
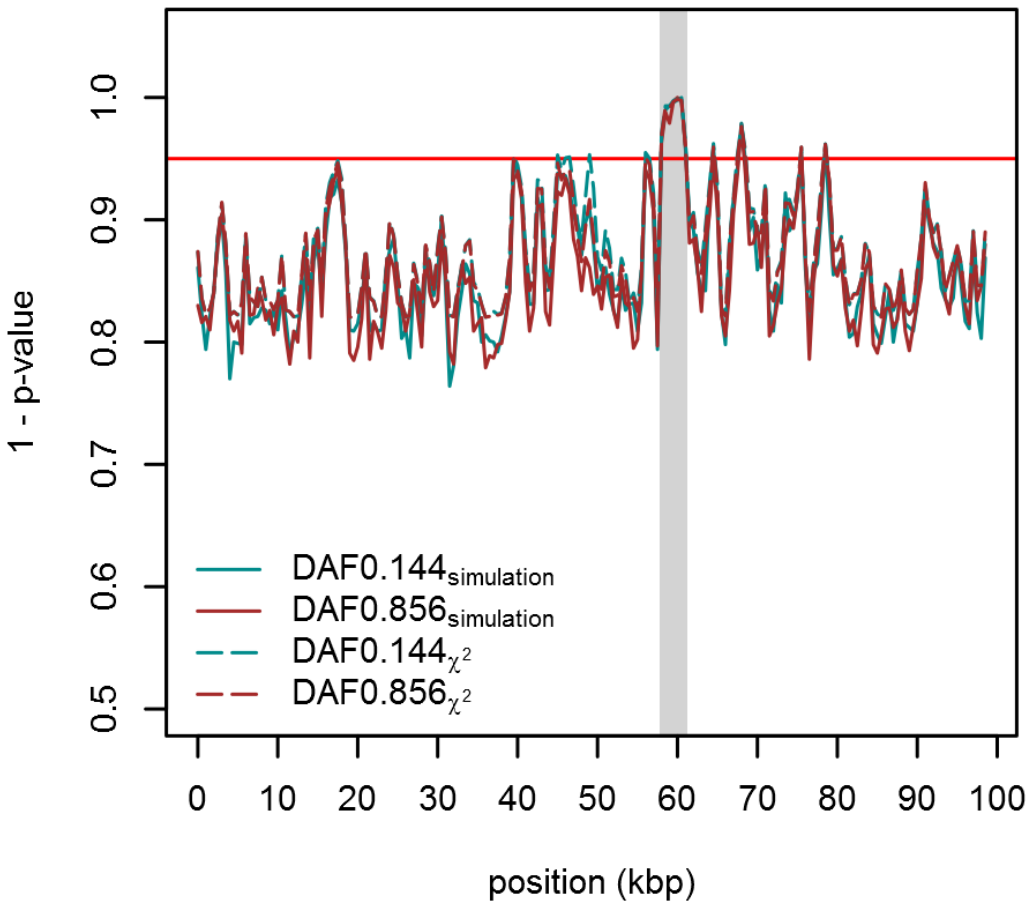


Figure 5. A sliding window implementation of the modified HKA test, taking into account the observed sampling scheme, allele frequencies and recombination rate. One minus the p-value at each window for derived allele frequencies of 0.144 (blue lines) and 0.856 (red lines), using the chi-squared distribution (dashed lines) and simulations (solid lines). The red horizontal line indicates a p-value of 0.05. The shaded grey bar indicates the region for which divergence data at the *Red* locus is not available.



Supplementary Material

Maximum likelihood estimates of θ under equilibrium

For conducting simulations analogous to those which led to Figure 4, but assuming a constant population size, we estimated θ under equilibrium (no change in population size backwards in time) using data from the 24 reference loci. For the *i*-th locus, assuming complete linkage, Tavare (Tavaré 1984) has shown that the number of segregating sites, S_i, follows a probability distribution characterised by the following density function

$$P(S_i = s_i) = \frac{n_i - 1}{\theta L_i} \sum_{j=1}^{n_i - 1} (-1)^{j-1} \binom{n_i - 2}{j - 1} \left(\frac{\theta L_i}{j + \theta L_i}\right)^{s_i + 1} \quad (1)$$

where s_i , L_i and n_i are the observed number of segregating sites, the length (in bp) and the sample size at the i[th] reference locus, respectively. Given the observed data, $D = (s_1, s_2, \ldots, s_{24})$, at our reference loci, the ln-likelihood of θ can be expressed as

$$L(D|\theta) = \sum_{i=1}^{24} \ln[P(S_i = s_i)] \quad (2)$$

Our maximum likelihood estimate of θ for the equilibrium model was $3.875*10^{-3}$ bp^{-1} (95% confidence intervals: 0.003107, 0.004705).

Estimating ρ under equilibrium

The following procedures were used to obtain ρ values that were likely to be compatible with data at the *Red* locus. We first simulated a neutral locus equivalent to the *Red* locus in length using SelSim (Spencer and Coop 2004), conditioning on our sampling scheme (six each of *b* and *R* alleles, from a wild allele frequency distribution of 0.856 and 0.144, with each allele being treated as derived in turn). To generate neutral variants, the simulations were conducted either conditioning on the observed number of segregating sites at the *Red* locus or using the maximum-likelihood estimates of θ from our 24 reference loci (see Supplementary Material - Maximum likelihood estimates of θ under

equilibrium). A range of values of ρ, from 0 to 0.005bp^{-1}, were considered. We performed 1000 simulations for each combination of parameters. We then estimated the R_M on each simulated dataset and obtained the mean across replicates. The average values of R_M were plotted against ρ, and ρ values with average R_M comparable to the observed R_M were regarded as compatible with the data.

Results

We estimated the minimum parsimonious number of recombination events (R_M) at the *Red* locus as 58. By simulating a region equivalent in length, sample size and allele frequencies, and using a range of values of ρ, we inferred that a value of R_M = 58 is compatible with $\rho \approx$ 0.0015-0.002 under the equilibrium situation (i.e. with no change in population size) (Figure S2). The relationship between R_M and ρ was reasonably insensitive to both the derived allele frequency (DAF) and to whether we conditioned on the number and locations of SNPs in the *Red* locus, or on the MLE estimate of θ from our reference loci (Figure S2). We assumed a conservative value of ρ = 0.001 in subsequent analyses under the equilibrium model. This estimate lies within the estimates of ρ bp^{-1} at the control loci (Table S3).

Coalescent simulations of the Red locus – under a constant population size

In addition to our simulations which incorporated a population size change, we simulated a 99669bp region (equivalent in length to the *Red* locus) under an equilibrium situation using SelSim (Spencer and Coop 2004), which uses the coalescent process to generate neutral polymorphism data, conditioning on the allele frequency trajectory of a single linked site (Griffiths 2003; Coop and Griffiths 2004). As with our population size change simulations using mbs, we placed the focal site (see Figure 1A) in the centre of the simulated *Red* locus. To some extent, we also tested the effect of placing the focal site in different places within the core locus and in order to test the robustness of our results. In most cases, we set the value of ρ as 0.001 (see Supplementary Material – Estimating ρ under equilibrium).

As well as incorporating uncertainty surrounding the derived allele frequency (DAF) (as with the population size change simulations), we also incorporated uncertainty surrounding θ in these simulations. We drew values of DAF from the posterior

distribution of allele frequencies, which is proportional to equation (1) in the main text. We drew DAFs in the same way as under mbs (see Methods in the main text). We used the same scheme to draw values of the population mutation rate, θ, making use of equation (15) in the Supplementary Material. For θ, we only consider values within the interval [1.7083e-3, 8.5015e-3], which correspond to the upper and lower bounds of θ, 30 ln-likelihood units below the maximum-likelihood estimate of 3.8751e-3bp^{-1}, see Supplementary Material - Maximum likelihood estimates of θ under equilibrium).

As with the population size change simulations, we replicated the missing data in our observed dataset by masking the same sites in our simulated datasets before calculating summary statistics. We carried out two sets of simulations, treating each allele (*R* and *b*) as derived in turn.

Results

For a DAF of 0.144, the observed data at the *Red* locus lies outside the 95% confidence intervals (CIs) of our coalescent simulations for many windows, for a range of summary statistics which incorporate both allelic classes (F_{ST}, D_{XY}, π_{total}, D_{total}) (Figure S5). The observed data generally do not lie outside the 95% CIs for within-allelic class summary statistics (D_{red}, D_{black}, π_{red}, π_{black}) (Figure S5). These results are qualitatively identical for a DAF of 0.856 (Figure S6).

<u>The classic HKA test using pi</u>

We carried out the classic HKA test of neutral evolution (Hudson et al. 1987) using custom scripts in R, and used ms (Hudson 2002) to test the significance of our results using the simulation method. As well as the test based on the number of segregating sites, we implemented a test based on the number of pairwise differences (π). For the case with polymorphism data from one species (as in our data) the equivalent of equations 1-5 from Hudson et al. (1987) using the number of pairwise differences as opposed to the number of segregating sites are as follows

$$E(\pi_i) = \theta_i \quad (3)$$

$$Var(\pi_i) = b_1\theta_i + b_2{\theta_i}^2 \tag{4}$$

$$E(D_i) = \theta_i(T+1) \tag{5}$$

$$Var(D_i) = E(D_i) + {\theta_i}^2 \tag{6}$$

$$\sum_{i=1}^{M} \pi_i = \sum_{i=1}^{M} \hat{\theta}_i,$$

$$\sum_{i=1}^{M} D_i = \left(\hat{T}+1\right)\sum_{i=1}^{M} \hat{\theta}_i, \tag{7}$$

$$D_i + \pi_i = \hat{\theta}_i\{T+1+1\}, \qquad i = 1, \ldots, M-1.$$

where

$$b_1 = \frac{n+1}{3(n-1)}$$

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$$

(Tajima 1989a). Solving system of equations (7) provides estimates of θ_i and T, which can be used to calculate Hudson et al. (1987)'s test statistic X^2:

$$X^2 = \sum_{i=1}^{M} \left(\pi_i - E(\pi_i)\right)^2 / Var(\pi_i) + \sum_{i=1}^{M} \left(D_i - E(D_i)\right)^2 / Var(D_i) \tag{8}$$

As with our summary of polymorphism, we implemented the HKA test in a sliding window along the *Red* locus. To do this, we took polymorphism and divergence data

from all 24 reference loci and one 1000bp *Red* locus window at a time, calculated Hudson et al. (1987)'s X^2 statistic using both S and π, and calculated p-values using the χ^2 distribution as well as by the simulation method described in Hudson et al. (1987). For the ~2.5kbp portion of the *Red* locus for which we do not have divergence data (the TE insertion which is not present in the Zebra finch lineage), we calculated divergence using 3kbp (1.5kbp from each side) of the *Red* locus immediately adjacent to the missing data.

Our sliding window implementation of the classic HKA test for S and π returned significant results for multiple windows across the *Red* locus, particularly around the region of the TE insertion which does not exist in the Zebra finch lineage (Figure S11).

The variance of the p-values calculated using π is generally larger than using S (Figure S11), using either the χ^2 distribution or simulations. The trends of the two measures of polymorphism between windows are similar, however. In general, for π, the simulation method returns more significant values than the χ^2 distribution. The converse is true for S. For the windows with the smallest p-values, π tends to return more significant results than S, but the discrepancy between the two is seems to be small.

The classic HKA and unequal locus lengths

The test statistic X^2 of Hudson et al. (1987) is expected to be approximately chi-squared (χ^2) distributed if certain assumptions are met. These are that S_i (or π_i) and D_i are stochastically independent of each other, and are normally distributed. These assumptions are thought to be valid if the samples sizes and the estimate of T are large enough, and if the loci under consideration are unlinked (Hudson et al. 1987).

We were interested in testing whether this was the case with unequal locus lengths, as in our case if we take the *Red* locus as a single non-recombining region (i.e. we do not use a sliding window approach) and compare it with our 24 (shorter) reference loci. To do so, we simulated two unlinked loci with a sample size of 12, a known length and θ, under the standard neutral model using ms (Hudson 2002), and then carried out the standard HKA test procedure (using the number of segregating sites) on the simulated polymorphism data, 1000 times. We plotted the resulting distribution of p-values assuming that the X^2 statistic is χ^2 distributed (Figure S3), as well as by using the simulation method described in (Hudson et al. 1987). We found that the X^2 statistic is

not χ^2 distributed in the case where either the two loci under consideration were of different sizes to each other, or if they had very different population mutation rates (per basepair). In these cases, a deviation from parity between the two loci results in the distribution of p-values being anti-conservative (Figure S3). The simulation method returned the expected distribution of p-values (data not shown).

The classic HKA and a change in population size

We were also interested in testing whether a change in population size, like the one that we inferred from our Z-linked intronic data, would affect the HKA test. To do so, we followed a similar procedure to the section above. We simulated two unlinked loci with a sample size of 12, a known length and θ, under the standard neutral model using ms (Hudson 2002), but incorporating a demographic expansion equivalent to that which we inferred from our data. We then carried out the standard HKA test procedure (using the number of segregating sites) on the simulated polymorphism data, 1000 times. We plotted the resulting distribution of p-values assuming that the X^2 statistic is χ^2 distributed (Figure S4). We found that the X^2 statistic is not χ^2 distributed in the case of a change in population size, but that the distribution of p-values is rendered conservative (Figure S4).

Supplementary Table S1. Allele counts at the *Red* locus from 161 wild-caught genotyped birds.

	Genotype		
Sex	*R*/*R* or *R*/-	*R*/*b*	*b*/*b* or *b*/-
Male	0	27	62
Female	9	n/a	63

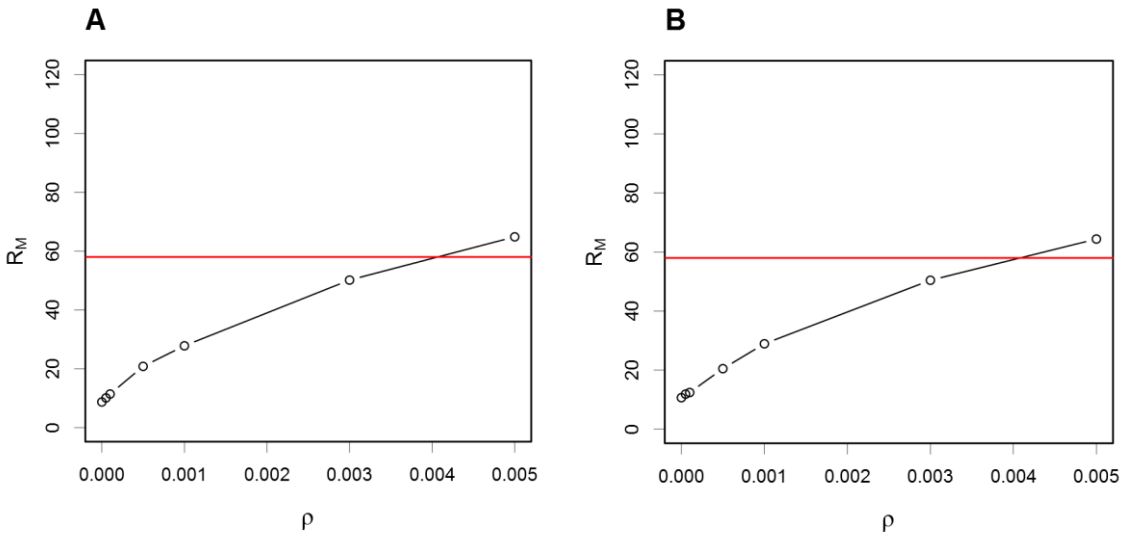
Supplementary Table S2. Testing R_M with known ρ.

ρ (bp^{-1})	Replicate	R_M	
		DAF = 0.1	DAF = 0.9
0	1	0	0
	2	0	0
	3	0	0
	4	0	0
	5	0	0
	6	0	0
	7	0	0
	8	0	0
	9	0	0
	10	0	0
0.0002	1	5	6
	2	3	12
	3	5	10
	4	7	9
	5	8	4
	6	14	4
	7	9	5
	8	2	7
	9	9	5
	10	8	10
0.002	1	34	42
	2	48	38
	3	36	51
	4	37	49
	5	41	36
	6	38	37
	7	41	38
	8	44	46
	9	38	38
	10	35	32

Supplementary Table S3. Recombination statistics for 24 Z-linked intronic reference loci. Values of ρ were obtained using LDhat (Auton and McVean 2007). Values of R_M were obtained using RecMin (Myers and Griffiths 2003).

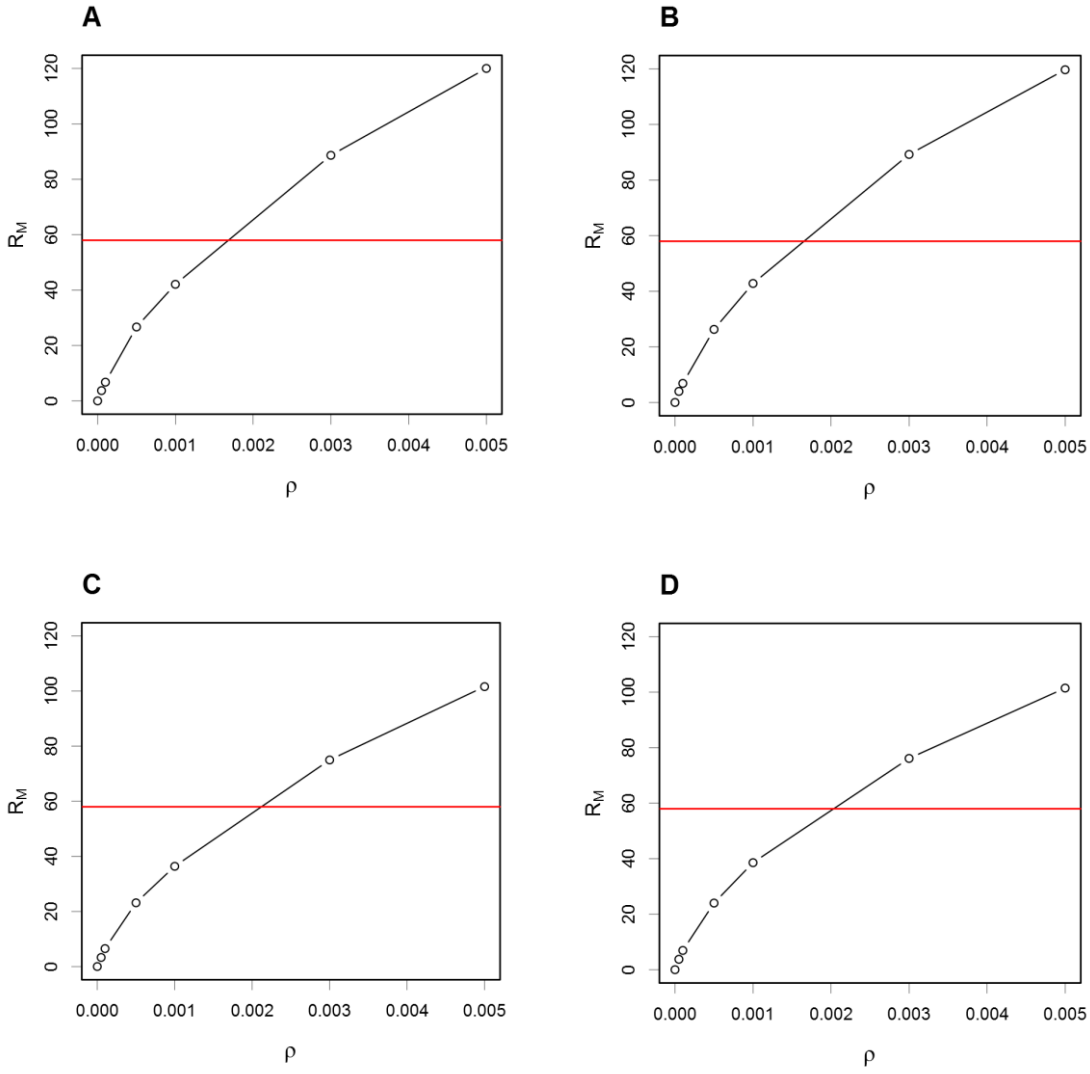
Locus	Length (bp)	ρ (bp^{-1})	R_M (locus^{-1})
KATNAL2_87636	1076	0.019516729	2
LPL_4548216	1156	0.015570934	3
ADAMTS19_5276	965	0.071502591	5
CNTNAP4_84481	1075	0.04372093	2
DAPK1_1020973	972	0.030864198	0
PPIP5K2_17622	1043	0	0
NUDT12_17751596	970	0.050515464	0
SNX2_23302562	1078	0	0
RASA1_2456501	1160	0.098275862	0
TRIM36_271262	1056	0	0
TSTD2_3165523	952	0.003151261	1
PIK3C3_348486	953	0.002098636	0
MTMR12_404680	995	0.00201005	0
LIFR_42671503	1090	0	0
DDX4_47480505	1050	0.005714286	1
ERCC8_49384755	1123	0.001780944	0
CWC27_5058075	931	0.005370569	1
RASEF_5301000	1049	0.000953289	0
FOCAD_5770866	1055	0	0
SNAPC3_599286	1054	0	0
similar_to_MP	1033	0	0
CDC37L1_64136	880	0.027272727	0
DOCK8_6560464	866	0.023094688	0
DHFR_72301206	978	0.040899796	2

Supplementary Figure S1



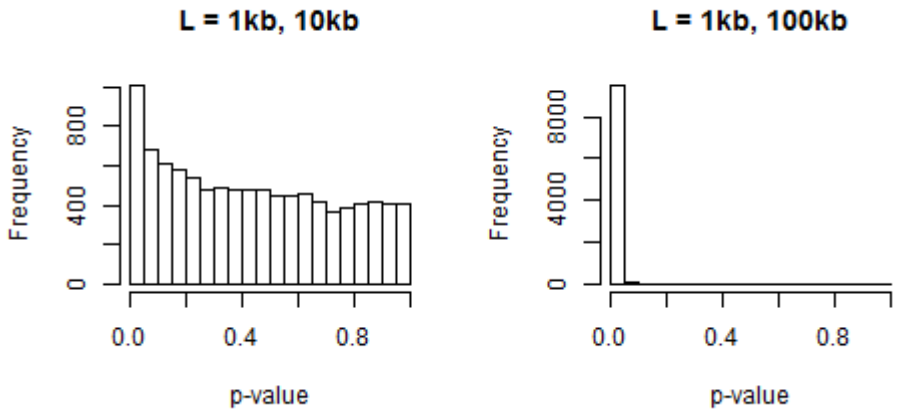
Legend. Comparison of the observed value of R_M at the *Red* locus (R_M = 58, red horizontal line) with values of R_M obtained from a simulated representation of the *Red* locus, for derived allele frequencies of 0.144 (panel A) and 0.856 (panel B), conditioning on a value of $\theta = 1.3458*10^{-2}$, derived from 24 Z-linked intronic reference loci and incorporating a change in population size.

Supplementary Figure S2



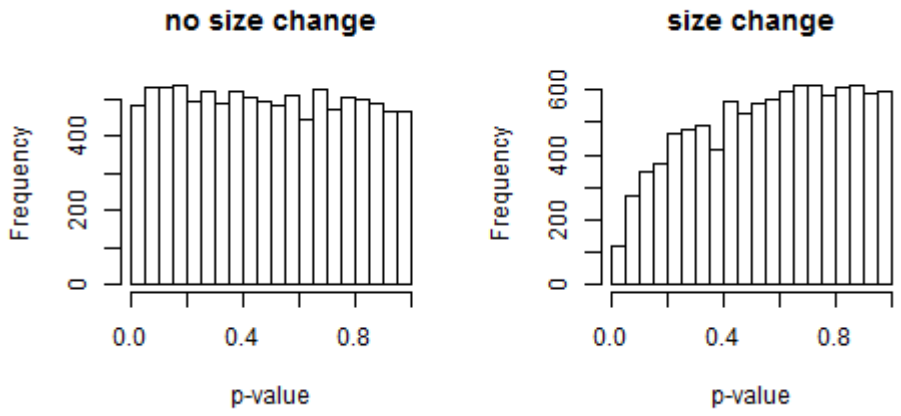
Legend. Comparison of the observed value of R_M at the *Red* locus (R_M = 58, red horizontal line) with values of R_M obtained from a simulated representation of the *Red* locus under equilibrium (no change in population size), for derived allele frequencies of 0.144 (panels A and C) and 0.856 (panels B and D), and either conditioning on the observed number of segregating sites at the *Red* locus (panels A and B) or the ML value of $\theta = 0.3875*10^{-3}$, derived from the 24 Z-linked intronic reference loci (panels C and D).

Supplementary Figure S3



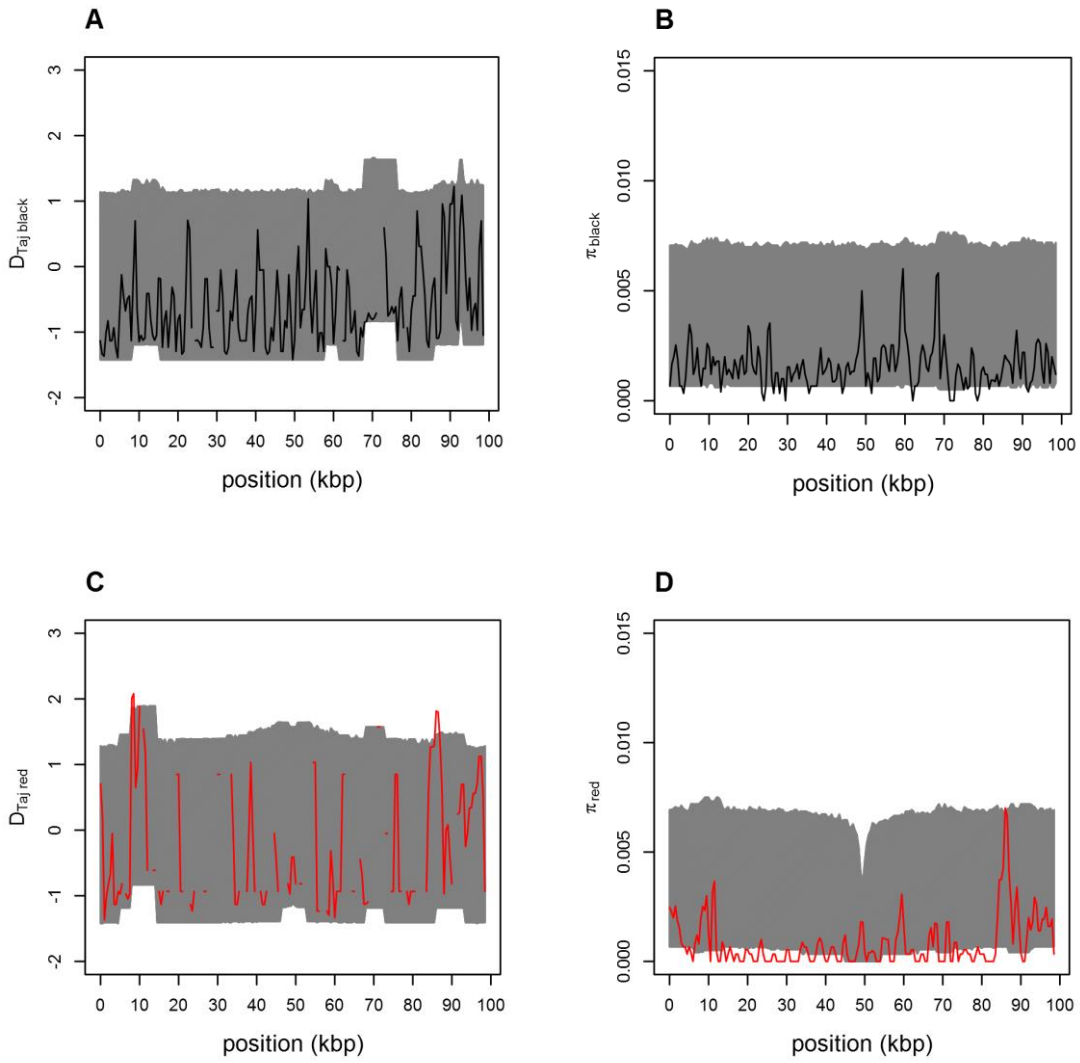
Legend. The effect of unequal locus lengths on the HKA test's *p*-value distribution, which were obtained from a χ^2 distribution with 1 d.f. We simulated two loci with 12 chromosomes under the standard neutral model, with values of θ for polymorphism and divergence informed by the Gouldian finch – Zebra finch system, either 1 kbp and 10kbp in length, or 1 kbp and 100kbp in length. If data are obtained with equal sized loci the *p*-values should be uniformly distributed, as shown in the left-hand panel of Supplementary Figure S4 (below). The (greater) surfeit of low *p*-values with (more) unequal locus lengths suggests the test is be anti-conservative when locus lengths are unequal.

Supplementary Figure S4



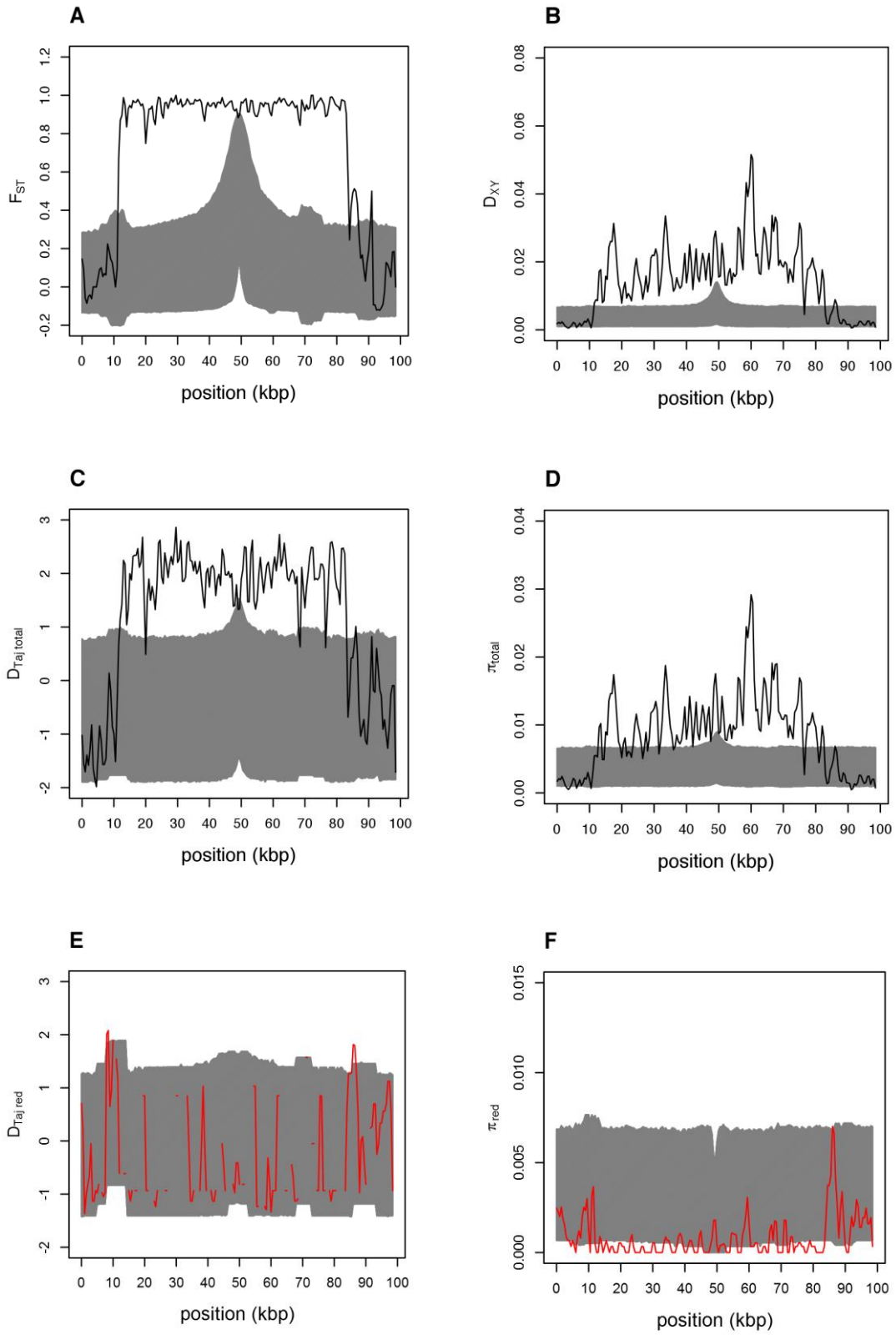
Legend. The effect of a population size change on the HKA test's *p*-value distribution, which were obtained from a χ^2 distribution with 1 d.f. Under no size change, we simulated two 1kbp loci with 12 chromosomes under the standard neutral model, with values of θ for polymorphism and divergence informed by the Gouldian finch – Zebra finch system. The size change model is identical, with the exception that we included the inferred population expansion in the Gouldian finch lineage in our simulations. If data are obtained from the model assumed in the original formulation of the HKA test, the *p*-values should be uniformly distributed, as shown in the left-hand plot. The deficit of low *p*-values under a size change model (right-hand plot) suggests the test should be conservative.

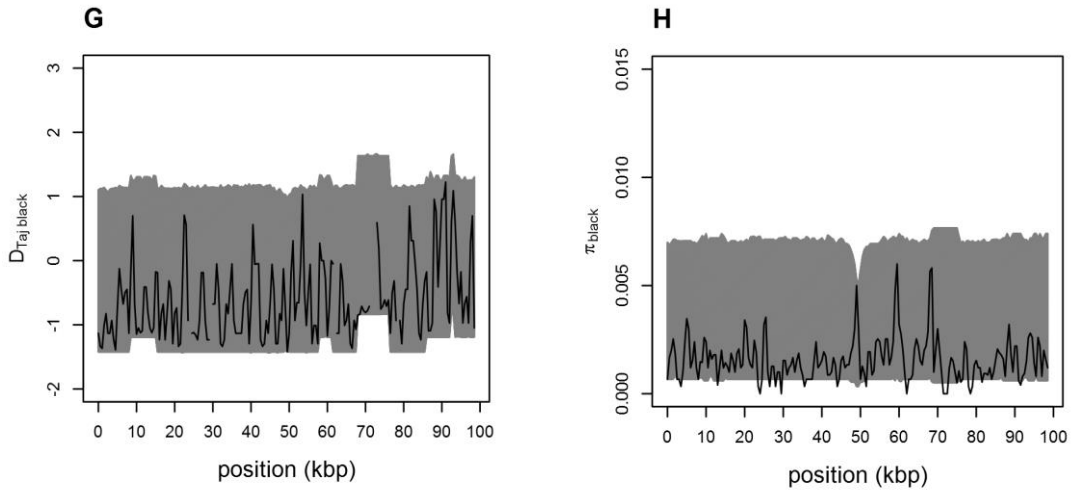
Supplementary Figure S5



Legend. A sliding window comparison of observed within-allelic class polymorphism at the *Red* locus and polymorphism simulated under the standard neutral model of molecular evolution incorporating a population size change, with a derived allele frequency of 0.144, taking into account the observed sampling scheme and recombination rate. Solid lines represent the observed data. The grey shaded region represents the space encompassed by the 95% confidence intervals derived from the simulations. **A**, Tajima's D within the black (ancestral) allelic class; **B**, nucleotide diversity within the black allelic class; **C**, Tajima's D within the red (derived) allelic class; **D**, nucleotide diversity within the red allelic class.

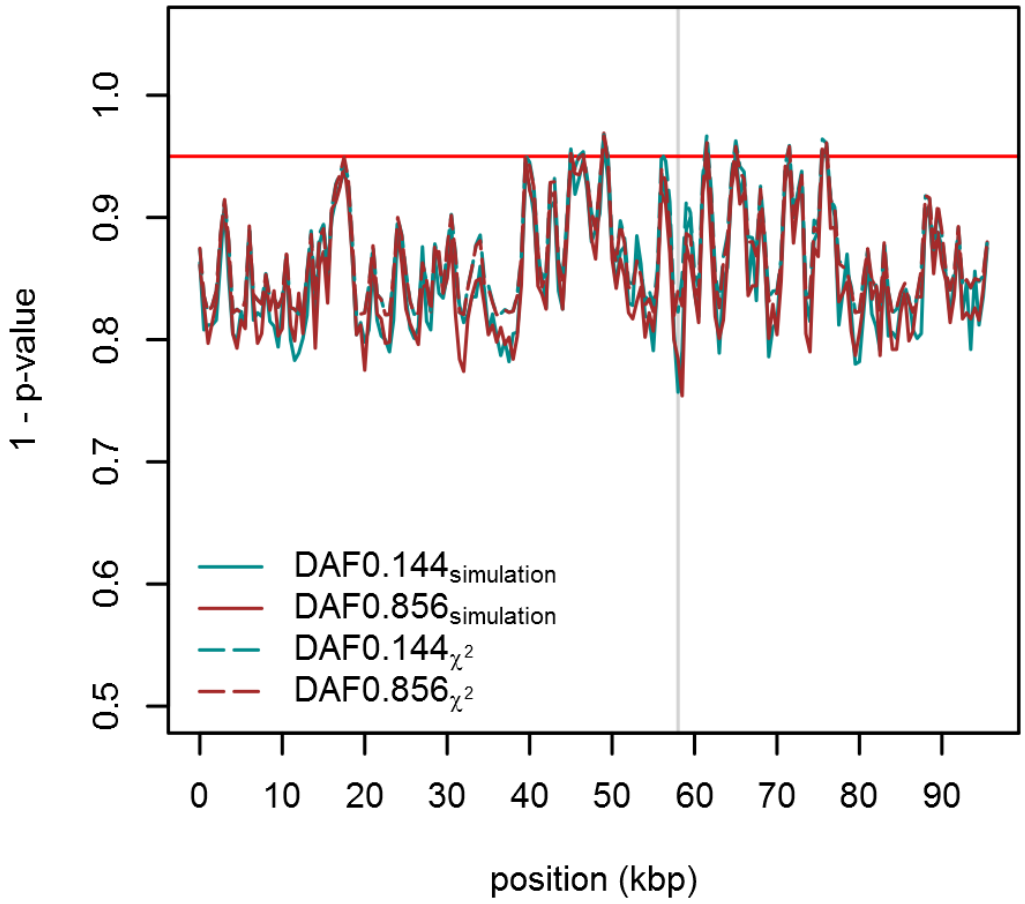
Supplementary Figure S6





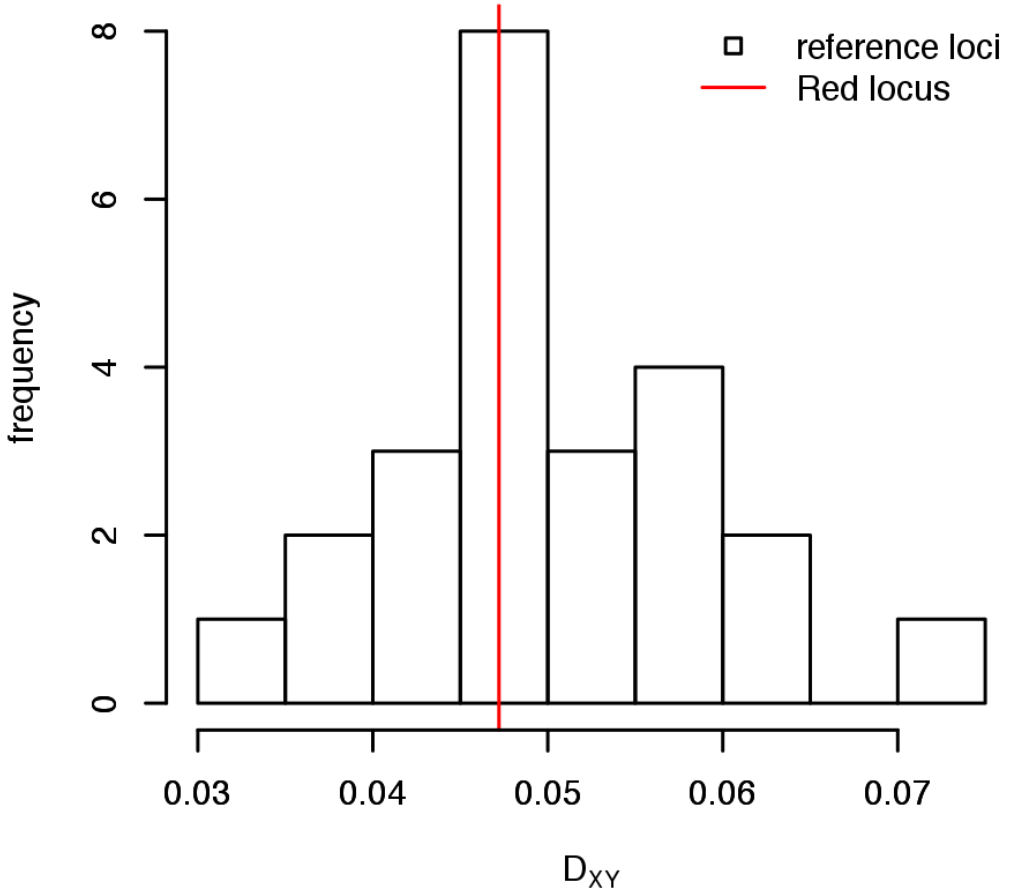
Legend. A sliding window comparison of observed between- and within-allelic class polymorphism at the *Red* locus and polymorphism simulated under the standard neutral model of molecular evolution incorporating a population size change, with a derived allele frequency of 0.856, taking into account the observed sampling scheme and recombination rate. Solid lines represent the observed data. The grey shaded region represents the space encompassed by the 95% confidence intervals derived from the simulations. **A**, F_{ST} between both allelic classes at the *Red* locus; **B**, D_{XY} between both allelic classes; **C**, Tajima's D for both allelic classes combined; **D**, nucleotide diversity for both allelic classes combined; **E**, Tajima's D within the red (ancestral) allelic class; **F**, nucleotide diversity within the red allelic class; **G**, Tajima's D within the black (derived) allelic class; **H**, nucleotide diversity within the black allelic class.

Supplementary Figure S7



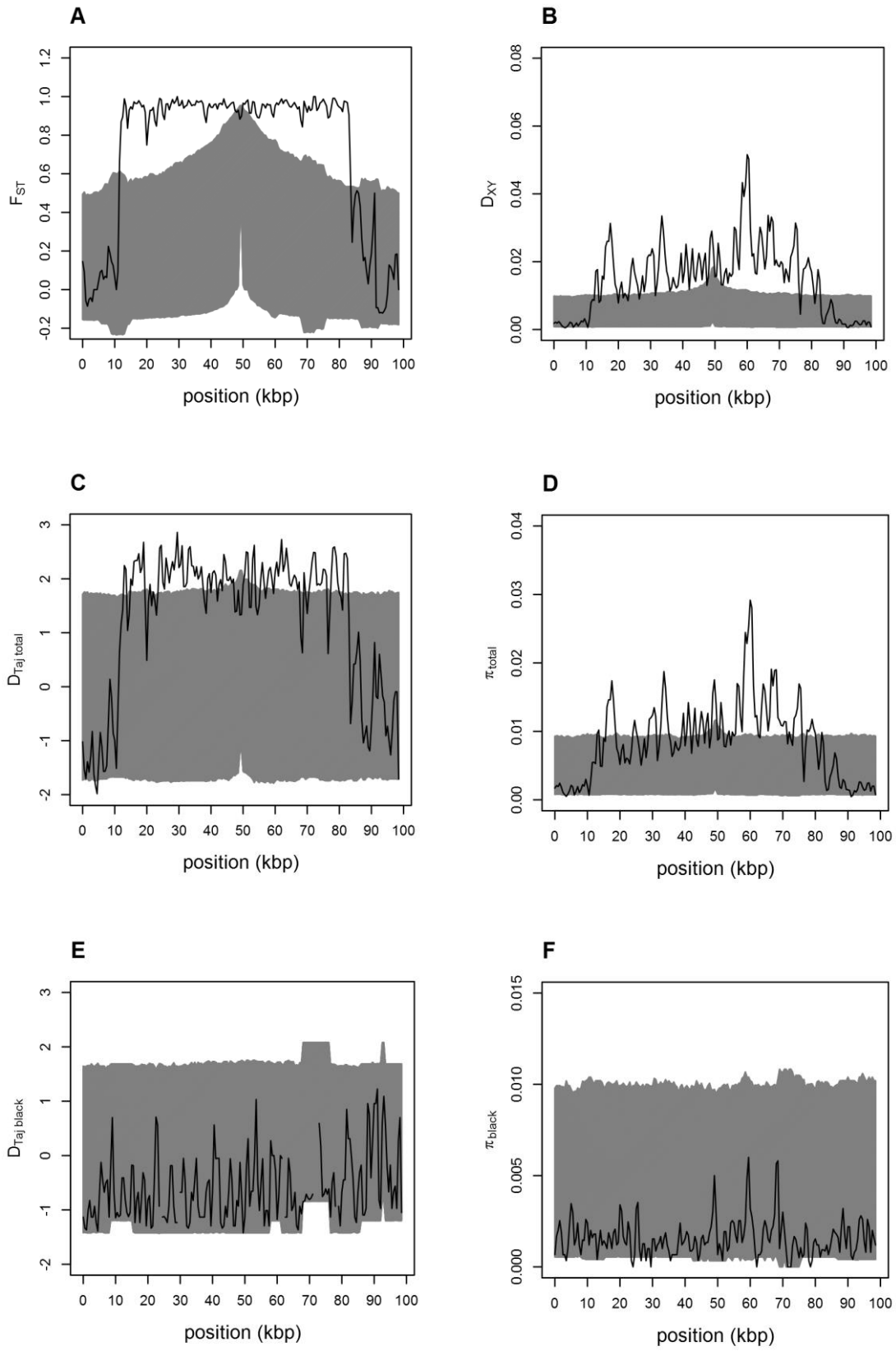
Legend. Sliding window implementation of the HKA test, taking into account non-random sampling and recombination. The region of the *Red* locus incorporating the putative transposable element in the Gouldian Finch lineage has been removed, and the position of the focal site set to its position (the vertical grey line). The horizontal red line indicates a p-value of 0.05.

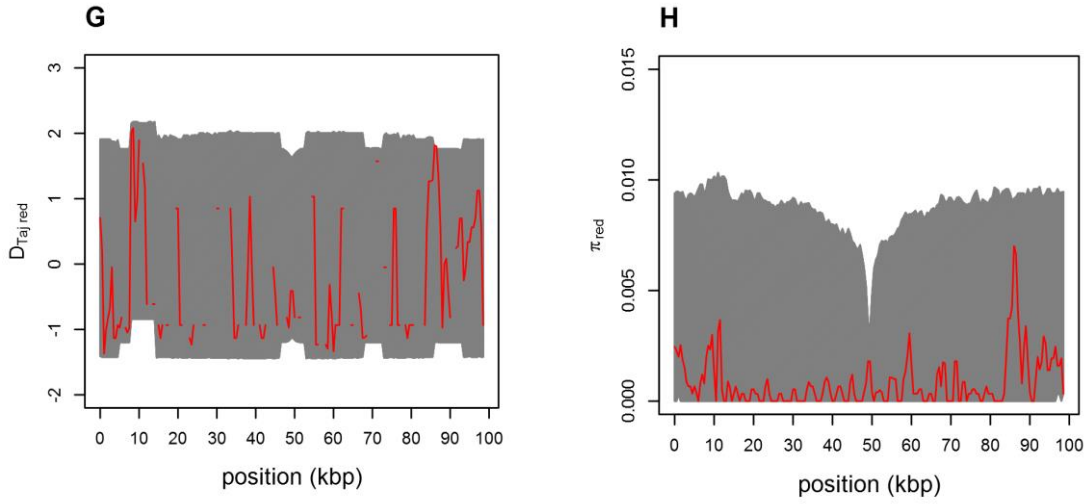
Supplementary Figure S8



Legend. The distribution of pairwise distance (D_{XY}) between Gouldian Finch and Zebra Finch at the 24 Z-linked reference loci (histogram bars) compared to D_{XY} between Gouldian Finch and Zebra Finch at the *Red* locus (red vertical line).

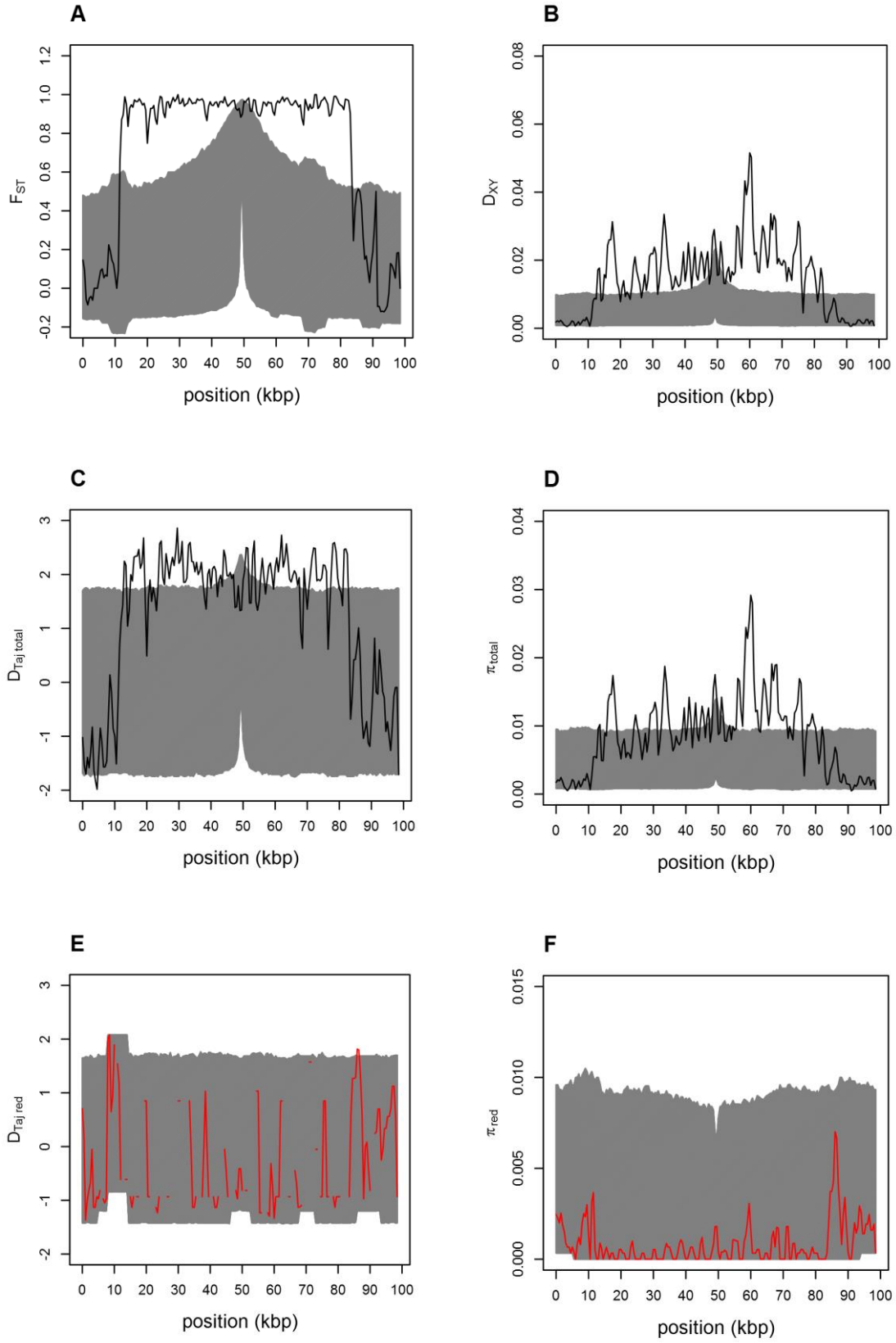
Supplementary Figure S9

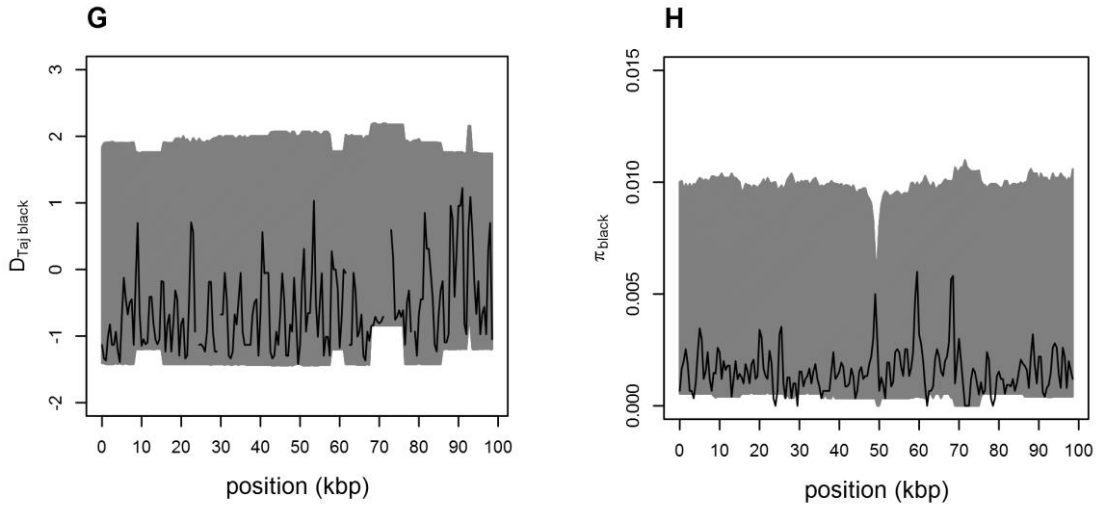




Legend. A sliding window comparison of observed polymorphism at the *Red* locus and polymorphism simulated under equilibrium and the standard neutral model of molecular evolution with a derived allele frequency of 0.144, taking into account the observed sampling scheme and recombination rate. Solid lines represent the observed data. The grey shaded region represents the space encompassed by the 95% confidence intervals derived from the simulations. **A**, F_{ST} between both allelic classes at the *Red* locus; **B**, D_{XY} between both allelic classes; **C**, Tajima's D for both allelic classes combined; **D**, nucleotide diversity for both allelic classes combined; **E**, Tajima's D within the black (ancestral) allelic class; **F**, nucleotide diversity within the black allelic class; **G**, Tajima's D within the red (derived) allelic class; **H**, nucleotide diversity within the red allelic class.

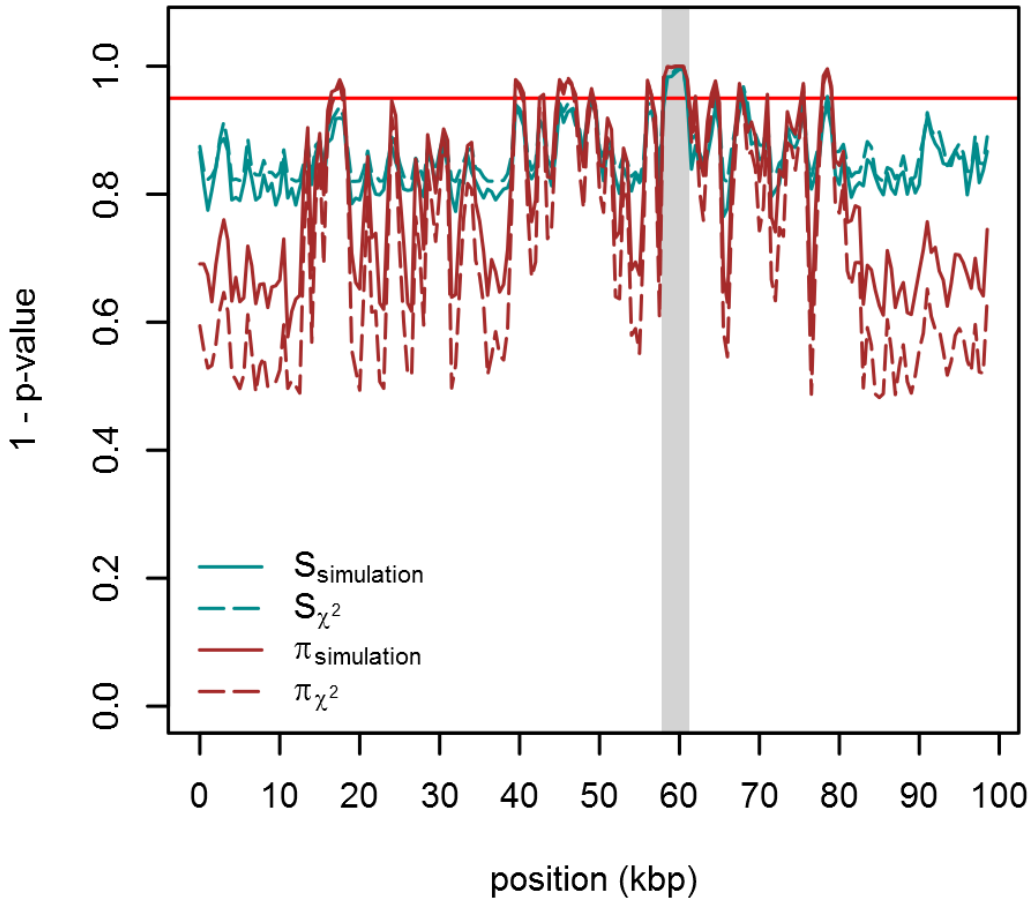
Supplementary Figure S10





Legend. A sliding window comparison of observed polymorphism at the *Red* locus and polymorphism simulated under equilibrium and the standard neutral model of molecular evolution with a derived allele frequency of 0.856, taking into account the observed sampling scheme and recombination rate. Solid lines represent the observed data. The grey shaded region represents the space encompassed by the 95% confidence intervals derived from the simulations. **A**, F_{ST} between both allelic classes at the *Red* locus; **B**, D_{XY} between both allelic classes; **C**, Tajima's D for both allelic classes combined; **D**, nucleotide diversity for both allelic classes combined; **E**, Tajima's D within the red (ancestral) allelic class; **F**, nucleotide diversity within the red allelic class; **G**, Tajima's D within the black (derived) allelic class; **H**, nucleotide diversity within the black allelic class.

Supplementary Figure S11



Legend. A sliding window implementation of the classic HKA test, using the number of segregating sites (blue lines) and nucleotide diversity (red lines), with significance testing using the chi-squared distribution (dashed lines) and simulations (solid lines). The red horizontal line indicates a p-value of 0.05. The shaded grey bar indicates the region for which divergence data at the *Red* locus is not available, due to the insertion of a putative transposable element in the Gouldian Finch lineage.

Appendix to Chapter 3

Solving the system of equations underlying the HKA test

Eq (5) of Hudson et al (1987) for the case when we have polymorphism data from only one species are

$$\sum_{i=1}^{M} s_i = \sum_{i=1}^{M} C(n_i)\hat{\theta}_i, \tag{1.1}$$

$$\sum_{i=1}^{M} D_i = \left(\hat{T} + 1\right) \sum_{i=1}^{M} \hat{\theta}_i, \tag{1.2}$$

$$s_i + D_i = \hat{\theta}_i\{\hat{T} + 1 + C(n_i)\}, \qquad i = 1, \dots, M - 1. \tag{1.3}$$

In Eq (1), M is the number of loci; s_i and D_i are the number of segregating sites and the divergence between two randomly chosen alleles at the i^{th} locus, respectively and $C(n_i) = \sum_{j=1}^{n-1} 1/j$, where n is the sample size at the i^{th} locus. These parameters are observed from data. $\hat{T}$ is the divergence time in scaled by $2N_e$ generations between ingroup and outgroup, and $\hat{\theta}_i$ is the population mutation rate (Watterson 1975) at the i^{th} locus. In order to calculate the X^2 statistic of Hudson et al (1987), $\hat{T}$ and $\hat{\theta}_i$ must be estimated by solving the system of $M + 1$ equations (1).

In order to solve Eq (1), we first want to incorporate the possibility of unequal lengths for the polymorphism and divergence comparisons at each locus. After Hudson et al (1987), we can rewrite Eq (1) defining $\hat{\theta}_i$ *per site*. To do so, we define two new parameters: l_{si} and l_{Di}, which are the length of the polymorphism comparison and the divergence comparison at locus i, respectively.

$$\sum_{i=1}^{M} s_i = \sum_{i=1}^{M} C(n_i) l_{s_i} \hat{\theta}_i, \tag{2.1}$$

$$\sum_{i=1}^{M} D_i = \left(\hat{T} + 1\right) \sum_{i=1}^{M} l_{D_i} \hat{\theta}_i, \tag{2.2}$$

$$s_i + D_i = l_{D_i} \hat{\theta}_i \left(\hat{T} + 1\right) + l_{s_i} \theta_i C(n_i), \qquad i = 1, \ldots, M-1. \tag{2.3}$$

First, we can express $\hat{\theta}_i, i = 1, \ldots, M-1$, each in terms of $\hat{T}$, by rearranging Eq (2.3):

$$\hat{\theta}_i = \frac{s_i + D_i}{l_{D_i}\left(\hat{T} + 1\right) + l_{s_i} C(n_i)}, \qquad i = 1, \ldots, M-1. \tag{3}$$

We can also express $\hat{\theta}_M$ in terms of $\hat{T}$ by substituting Eq (3) into Eq (2.1), and rearranging:

$$\hat{\theta}_M = \frac{\sum_{i=1}^{M} s_i - \sum_{i=1}^{M-1} \left(C(n_i) l_{s_i} . \frac{s_i + D_i}{l_{D_i}\left(\hat{T} + 1\right) + l_{s_i} C(n_i)} \right)}{C(n_M) l_{s_M}}, \tag{4}$$

As we now have expressions for each $\hat{\theta}_i, i = 1, \ldots, M$, in terms of $\hat{T}$, we can substitute the RHS of Eq (3) and Eq (4) into Eq (2.2) and express $\hat{T}$ entirely in terms of known quantities ($f\left(\hat{T}\right)$ below), and so we can solve for $\hat{T}$. We do this using the uniroot() function in the R language (R Core Team 2014), after constructing $f\left(\hat{T}\right)$ programmatically (see below).

Given our estimate of $\hat{T}$, we can solve Eq (2.3) for $\hat{\theta}_i, i = 1, \ldots, M-1$, and given these estimates of $\hat{\theta}_i, i = 1, \ldots, M-1$ we can solve either Eq (2.1) or Eq (2.2) for $\hat{\theta}_M$. We now have an estimate of theta for each locus ($\hat{\theta}_i$), and of the divergence time ($\hat{T}$).

We can now use Eqs (1-4) in Hudson et al (1987) to obtain the expectation and the variance of the number of segregating sites and the divergence at each locus, with which we can calculate Hudson et al (1987)'s X^2 statistic.

Computer code to perform the rearrangement and substitution described above, and to calculate the X^2 statistic, was written in R, and is presented below. This script works with an arbitrary number of loci, each of which may have different sample sizes. The

data from Hudson et al (1987) is presented at the end as an example, and can be used as a check.

For the case where we considered non-random sampling and recombination at the *Red* locus in Gouldian finches, the expectation and variance of the number of segregating sites at a window of the *Red* locus were derived from simulation, as opposed to the standard neutral model (see Eq (12) in Chapter 3). We modified our R code to incorporate Eq (12) in Chapter 3 for this purpose, but the code presented below is for the original formation of the HKA test (i.e Eq (2) of this Appendix).

```
# This script should solve system of equations 5 in Hudson, Kreitman and Aguade
# (1987). A test of neutral molecular evolution based on nucleotide data. Genetics
# 116: 153-159,
# for an arbitrary number of loci.
#
# Equation numbers below are as in the HKA paper
# I have split system of equations 5 into 5i, 5ii, 5iii and 5iv:
#
# 5i = Sum of segregating sites in species A
# 5ii = Sum of segregating sites in species B (NOT USED)
# 5iii = Sum of divergent sites
# 5iv = Si + Di for loci i,..,L-1 (there are L-1 equations for 5iv)
#
# The input table is in the following format:
#
#   myTable <- data.frame(D = c(210, 18),
#                         S = c(9, 8),
#                         lenD = c(4052, 324),
#                         lenS = c(414, 79),
#                         n = c(81, 81))
#
# With a column each for the number of divergent sites (D), the number of segregating
# sites (S), the length of the intra-specific comparison locus (lenS), and the length of
# the inter-specific comparison locus (lenD).

# ONE RESTRICTION:

# IT ONLY WORKS WITH POLYMORPHISM DATA FROM ONE SPECIES


#A function to make a function, needed to solve for Tplus1 at solve_step_1:
make_function <- function(args, body, env = parent.frame()) {
```

```
  as.function(c(args, body), env)
}

# The first step is to express theta_i...theta_L-1 in terms of
# T + 1 (I'm going to call 'T + 1', 'Tplus1'), using the last
# equations (i.e. all of 5iv)

# The general structure (for the i_th locus) is:

# D_i + S_i = (length_D_i * theta_i * Tplus1) + (CNa * length_S_i * theta_i)

# and rearranged, as we want it:

# theta_i = D_i + S_i / ((length_D_i * Tplus1) + (length_S_i * CNa))

#This function expresses each theta (except theta_L) in terms of Tplus1:
rearrange_step_1 <- function(D, S, lenD, lenS, CNa){
  myString <- paste0('(', D, ' + ', S, ') / ((', lenD, ' * Tplus1) + (', lenS, ' * ', CNa,'))')
  return(myString)
}

# Now express theta_L in terms of Tplus1, by subbing the
# equations from rearrangement step 1 into equation 5i
rearrange_step_2 <- function(myTable){
  n <- nrow(myTable) # n is the number of loci
  CNas <- sapply(myTable$n, FUN = function(x) sum(1/seq_len(x - 1)) )
  # for theta in i=1 to i=L-1, make a function body expressing
  # theta_i in terms of Tplus1, using rearrange_step_1:
  tempBodies <- vector('character', length = n-1)
  for(i in seq_len(n-1)){
    tempBodies[i] <- rearrange_step_1(D = myTable$D[i],
                                      S = myTable$S[i],
```

```
                                        lenD = myTable$lenD[i],
                                        lenS = myTable$lenS[i],
                                        CNa = CNas[i])
  }

  # Now substitute the above function bodies into a rearranged equation 5i,
  # and express theta_L (LHS) in terms of Tplus1 (RHS)
  tempVec <- vector('character', length = n-1)
  for(i in seq_along(tempVec)){
    tempVec[i] <- paste0(CNas[i],' * ', myTable$lenS[i], ' * ', '(', tempBodies[i], ')')
  }

  tempVec_2 <- paste0(tempVec, collapse = ' + ')

  theta_L <- paste0('(', sum(myTable$S), ' - ( ', tempVec_2, ')) / ( ', CNas[n], ' * ', myTable$lenS[n], ' )')

  return(theta_L) # this is equation 5i, rearranged with substitution,
  # expressing theta_L in terms of Tplus1
}

# now we can express equation 5ii entirely in terms of Tplus1, by substituting
# all thetas with expressions including only Tplus1, which we derived using
# rearrange_step_1 and rearrange_step_2
rearrange_step_3 <- function(myTable){
  n <- nrow(myTable) # n is the number of loci
  CNas <- sapply(myTable$n, FUN = function(x) sum(1/seq_len(x - 1)) )

  theta_L <- rearrange_step_2(myTable)

  theta_i <- vector('character', length = n-1)
  for(i in seq_len(n-1)){
    theta_i[i] <- rearrange_step_1(D = myTable$D[i],
```

```
                                    S = myTable$S[i],
                                    lenD = myTable$lenD[i],
                                    lenS = myTable$lenS[i],
                                    CNa = CNas[i])
  }

  lenDi_theta_i <- vector('character', length = n-1)
  for(i in seq_len(n-1)){
    lenDi_theta_i[i] <- paste0('(', myTable$lenD[i], ' * (', theta_i[i], '))')
  }

  lenD_L_theta_L <- paste0('(', myTable$lenD[n], ' * (', theta_L, '))')

  fun_Tplus1 <- paste0('((', paste0(lenDi_theta_i, collapse = ' + '),
                       ' + ', lenD_L_theta_L, ') * Tplus1) - ', sum(myTable$D))

  return(fun_Tplus1)
}


# Now run the above functions and solve for Tplus1:
solve_step_1 <- function(myTable){
  b <- parse(text = rearrange_step_3(myTable))[[1]]
  f <- make_function(args = alist(Tplus1 = NULL), body = b)
  Tplus1 <- uniroot(f, interval = c(0, 100))$root # may need to edit the interval
  return(Tplus1)
}

# NOW SOLVE FOR THETA_i..L-i, USING 5iv:
solve_step_2 <- function(myTable, Tplus1){
  n <- nrow(myTable) # n is the number of loci
  CNas <- sapply(myTable$n, FUN = function(x) sum(1/seq_len(x - 1)) )
```

```
  theta_i_val <- vector('numeric', length = n-1)
  for(i in seq_len(n-1)){
    theta_i_val[i] <- (myTable$D[i] + myTable$S[i]) / ((myTable$lenD[i] * Tplus1) + (CNas[i] * myTable$lenS[i]))
  }
  return(theta_i_val)
}


# NOW SOLVE FOR THETA_L, USING 5i
solve_step_3 <- function(myTable, theta_i){
  n <- nrow(myTable) # n is the number of loci
  CNas <- sapply(myTable$n, FUN = function(x) sum(1/seq_len(x - 1)) )
  theta_L <- (sum(myTable$S) - sum(myTable$lenS[1:(n-1)] * theta_i * CNas[1:(n-1)])) / (myTable$lenS[n] * CNas[n])
  return(theta_L)
}

# This function wraps everything above together:
HKA_solve_system_S <- function(myTable){

  Tplus1 <- solve_step_1(myTable)
  theta_i <- solve_step_2(myTable, Tplus1)
  theta_L <- solve_step_3(myTable, theta_i)

  return(list('t' = Tplus1 - 1, 'thetas' = c(theta_i, theta_L)))
}

# Calculate the X^2 statistic of Hudson, Kreitman and Aguad? (1987), A test of
# neutral molecular evolution based on nucleotide data. Genetics 116: 153-159
# (for the HKA test)
#
# E(S_i) = CNa * theta_i
# Var(S_i) = E(S_i) + (theta_i)^2 * (CNa)^2 <- not actually CNa-squared, but see formula
```

```r
#
#
# E(D_i) = theta_i * (T + 1)
#
# Var(D_i) = E(D_i) + theta_i^2
#
# Need to use my functions above to calculate estimates of theta and
# T + 1, to feed into the equations above, in order to get expected values and
# variances
#
# Then the test statistic is as follows:
#
# X^2 = sum( (S_i - E(S_i))^2 / Var(S_i) )
#     + sum( (D_i - E(D_i))^2 / Var(D_i) )

get_X_squared <- function(myTable){

  CNas <- sapply(myTable$n, FUN = function(x) sum(1/seq_len(x - 1)) )
  CNas_sq <- sapply(myTable$n, FUN = function(x) sum(1/(seq_len(x - 1))^2) )

  myEstimates <- HKA_solve_system_S(myTable)

  E_S <- myEstimates$thetas * CNas * myTable$lenS

  Var_S <- E_S + ((myEstimates$thetas * myTable$lenS)^2 * CNas_sq)

  E_D <- myEstimates$thetas * myTable$lenD * (myEstimates$t + 1)

  Var_D <- E_D + (myEstimates$thetas * myTable$lenD)^2

  X_2 <- sum((myTable$S - E_S)^2 / Var_S) +
    sum((myTable$D - E_D)^2 / Var_D)
```

```
  return(X_2)
}

#.....................................................................

# This is the input format:
# The example is taken from the original HKA paper (above)
HKAtable <- data.frame(D = c(210, 18),
                       S = c(9, 8),
                       lenD = c(4052, 324),
                       lenS = c(414, 79),
                       n = c(81, 81))

# Give the last function the table in the correct format to get the estimates of theta and T:
myEstimates <- HKA_solve_system_S(HKAtable)

# then you can get the X^2 stat of Hudson et al (1987):
X_squared <- get_X_squared(HKAtable)
# and the p-value:
pchisq(X_squared, df = 1, lower.tail = F)
```

Chapter 4. Evidence for ongoing selection for preferred codons in *Drosophila simulans*, with a comparison to *Drosophila melanogaster*

Contributing authors: Benjamin C. Jackson, José L. Campos, Brian Charlesworth and Kai Zeng

Abstract

We aim to identify the evolutionary forces affecting 4-fold degenerate sites in *Drosophila simulans* and *Drosophila melanogaster*, because these sites are often used to make inferences about selection and demography, and because they form an important component of the genome. Much previous work has tried to identify the extent to which selection for preferred codons occurs in *D. melanogaster*, but these analyses were complicated by an excess of GC → AT substitutions along the *D. melanogaster* lineage, which has led some workers to suggest that there is no ongoing selection for preferred codons in this species. Less is known about *D. simulans*, although there is evidence that its genome composition may be closer to equilibrium. We address these questions by using whole-genome polymorphism datasets from *D. simulans* and *D. melanogaster*. By using *D. yakuba* as an outgroup and a non-homogeneous substitution model that takes into account the effects of non-equilibrium base composition, combined with methods that infer selection from polymorphism data, we obtained evidence for ongoing selection for preferred codons in both species. Based on two methods that are able to infer selection intensity on long versus short timescales, respectively, we showed that, while selection for preferred codons has clearly become less effective in *D. melanogaster*, the level of selection in *D. simulans* has remained relatively stable.

Introduction

Here, we investigate the forces that affect evolution at 4-fold degenerate coding sites in *Drosophila simulans* and *D. melanogaster*. These sites represent a substantial compositional part of the genome, and are often used as references against which selection at other sites, for example non-synonymous sites, is tested (McDonald and Kreitman 1991; Rand and Kann 1996; Parsch et al. 2010; Stoletzki and Eyre-Walker 2011). Therefore, quantifying the forces that affect their evolution is also necessary for a general understanding of genome evolution and making robust inferences about demography and selection elsewhere in the genome (e.g. Matsumoto et al. 2016).

Codon usage bias (CUB) is a key feature of genomes. It is the disproportionate use of some codons among the set of codons that code for a single amino acid. There is evidence for CUB in a wide range of organisms, including both prokaryotes and eukaryotes (Drummond and Wilke 2008; Hershberg and Petrov 2008). The most common explanation for some codons being preferred is that this maximises translational efficiency and/or accuracy (Hershberg and Petrov 2008); avoiding the toxicity of misfolded proteins generated by ribosome errors has also been proposed as an explanation (Drummond and Wilke 2008). Recent work has also suggested the possibility that stabilizing, as opposed to directional, selection maintains the frequencies of synonymous codons (Charlesworth 2013; Fuller et al. 2014).

In most species of *Drosophila* for which data are available, including *D. melanogaster* and *D. simulans*, all the preferred codons are GC-ending (Vicario et al. 2007; Zeng 2010). Thus, selection for preferred codons acts to increase the GC content of third position sites in coding sequences, and GC-ending and AT-ending codons have been conventionally used as proxies for preferred and unpreferred codons, respectively. Because the preferred and unpreferred variants are so well characterised in this way, and because there are a large number of sites for which the genotype, and therefore the phenotype, is known, CUB should be a tractable quantitative trait to investigate. Further, because there are so many sites at which selection against unpreferred codons may act, it seems possible that the combined action of mildly deleterious mutations at many of these sites will have non-trivial implications for the amount of genetic load the population is subject to (reviewed in Chapter 4 of Charlesworth and Charlesworth 2010).

As in other species, evidence for selection for preferred codons in *D. melanogaster* comes from the fact that the level of codon bias is related to expression level (Table 3 in Campos et al. 2013) and that there is a negative relationship between the level of CUB and synonymous site divergence in the *Drosophila* subgroup, a pattern consistent with purifying selection maintaining optimal codons (Bierne and Eyre-Walker 2006). Lawrie et al (2013) also found evidence of strong purifying selection at 4-fold degenerate sites in *D. melanogaster*, but this force seems to be unrelated to CUB.

Previous studies, using a relatively small number of loci, compared *D. melanogaster* and *D. simulans* with respect to the extent of CUB. It was found that *D. melanogaster* has undergone a reduction in the strength of selection for preferred codons compared to *D. simulans* (Akashi 1995; Akashi 1996; McVean and Vieira 2001), but that selection for preferred codons is ongoing in *D. simulans* and may also be so, albeit much more weakly, in *D. melanogaster* (Kliman 1999).

Two main, non-exclusive explanations have been proposed for the observed reduction in CUB in *D. melanogaster*. These are firstly, that *D. melanogaster* has undergone a reduction in the population-scaled strength of selection for preferred codons, $4N_e s$, where N_e is the effective population size and s is the selection coefficient favouring preferred codons. This reduction in selection can be attributed either to a reduction in N_e (Akashi 1996), or a reduction in s, perhaps due to changing ecological conditions (Clemente and Vogl 2012a, 2012b). The second explanation is that *D. melanogaster* has undergone a shift in mutational bias towards AT alleles (Takano-Shimizu 2001; Kern and Begun 2005; Zeng and Charlesworth 2010a; Clemente and Vogl 2012b). It has also been argued that these two effects must be invoked together to explain the patterns in the *D. melanogaster* lineage (Nielsen et al. 2007; Clemente and Vogl 2012a, 2012b). Both of these can explain the observation that the *D. melanogaster* lineage has accumulated significantly more $\mathrm{GC} \rightarrow \mathrm{AT}$ substitutions than $\mathrm{AT} \rightarrow \mathrm{GC}$ ones since the last common ancestor of *D. melanogaster* and *D. simulans*, indicating that *D. melanogaster* is not currently at equilibrium with respect to GC content (Akashi 1995; Akashi 1996; Poh et al. 2012).

Irrespective of the reason(s) for non-equilibrium in *D. melanogaster*, it presents a problem for ancestral state reconstruction, a process that is necessary for inferring substitution patterns along a lineage of interest and for polarising segregating sites into ancestral and derived variants to understand their more recent evolution. Using

maximum parsimony methods or maximum likelihood models that assume equilibrium base composition under such circumstances can lead to erroneous inferences, although these two methods were adopted by many previous analyses in various *Drosophila* species (reviewed by Akashi et al. 2007; Matsumoto et al. 2015). Departures from base composition equilibrium may also lead to complex polymorphism patterns (Zeng and Charlesworth 2009), and both of these difficulties might explain the mixed evidence for nature of forces acting on synonymous sites in *D. melanogaster* (Zeng and Charlesworth 2010a; Clemente and Vogl 2012a).

One factor that may confound the study of CUB is GC-biased gene conversion (gBGC), which is a recombination-associated process, and acts to increase GC content at sites where recombination occurs (see Duret & Galtier 2009 for a review). Generally, studies have found little or no evidence for gBGC in *D. melanogaster* (Clemente & Vogl 2012b; Comeron et al. 2012; Campos et al. 2013; Robinson et al. 2014), although there is some evidence either for the action of selection for GC or gBGC on the evolution of non-coding sequences in *D. simulans* (Haddrill & Charlesworth 2008). Nevertheless, in order to control for gBGC, below we also analyse data from the 8-30bp region of short introns (SIs), which are widely considered to be otherwise neutrally evolving in *Drosophila* (Halligan & Keightley 2006; Parsch et al. 2010; Clemente & Vogl 2012b).

Overall, the evidence for selection for preferred codons in *D. melanogaster* is complicated by the non-equilibrium situation in this species. It seems important therefore to revisit this question using up-to-date methodology that explicitly takes into account non-equilibrium. Although there has been some suggestion that *D. simulans* is close to base composition equilibrium compared to *D. melanogaster* (Akashi 1995; Akashi 1996; McVean and Vieira 2001; Kern and Begun 2005; Akashi et al. 2006; Haddrill and Charlesworth 2008), much less is known about CUB in *D. simulans*. Here we aim to systematically compare the two species using whole genome polymorphism datasets. To this end, we acquire new whole genome data from *D. simulans* and use an existing genomic dataset for *D. melanogaster*. In addition, we employ methods that can infer selection intensity on different timescales, along the *D. melanogaster* and *D. simulans* lineages, with the aim of shedding further light on the evolutionary dynamics of genome composition in these two species.

Materials and Methods

Sequence data preparation

We first describe 22 new *D. simulans* isofemale lines, 11 of which were collected by William Ballard in 2002, from Madagascar (MD lines – MD03, MD146, MD197, MD201, MD224, MD225, MD235, MD238, MD243, MD255, MD72), and the other 11 were collected by P. Andolfatto in 2006, from Kenya (NS lines – NS11, NS111, NS116, NS19, NS37, NS49, NS63, NS64, NS89, NS95, NS96). We produced homozygous lines by full-sib inbreeding in the Charlesworth lab for nine generations; however, six lines (NS11, NS63, NS116, MD224, MD243, MD255) were lost early in the process of inbreeding. For these lines, we sequenced the initial stocks that we had received from the Andolfatto lab. Genomic DNA was prepared for each isofemale line by pooling twenty-five females, snap freezing them in liquid nitrogen, extracting DNA using a standard phenol-chloroform extraction protocol with ethanol, and ammonium acetate precipitation. These flies were sequenced by the Beijing Genomics Institute (BGI; http://bgi-international.com/). A 500bp short-insert library was constructed for each sample, and the final data provided consisted of 90bp paired-end Illumina sequencing (pipeline version 1.5), with an average coverage of 64X. We double-checked the quality of the filtered reads for each allele with FastQC (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and no further trimming was necessary. The raw reads were deposited in the European Nucleotide Archive, study accession number: PRJEB7673.

We obtained 20 further *D. simulans* isofemale lines from Rogers et al (2014). These lines were from the same sampling localities in Kenya (10 lines: NS05, NS113, NS137, NS33, NS39, NS40, NS50, NS67, NS78, NS79) and Madagascar (10 lines: MD06, MD105, MD106, MD15, MD199, MD221, MD233, MD251, MD63, MD73) as above. Each line was sequenced on between 2-3 lanes of paired-end Illumina sequencing at the UC Irvine High Throughput Genomics centre (http://ghtf.biochem.uci.edu/) per line. Further information about these lines and their sequencing is available in Rogers et al (2014). After examining FastQC files for these 20 lines, we trimmed two lines with apparently lower quality scores (MD233 and MD15) using the trim-fastq.pl script from popoolation 1.2.2 (Kofler et al. 2011) with the (minimum average per base quality score) --quality-threshold flag set to 20.

Downstream of sequencing, we combined both datasets and used a BWA/SAMtools/GATK pipeline, previously described in Campos et al (2014) and Jackson et al (2015), to generate genotype calls. Briefly, we aligned and mapped reads for each *D. simulans* line to the second generation assembly of the *D. simulans* reference sequence (Hu et al. 2013) using BWA 0.7.10 (Li and Durbin 2009). We used SAMtools 1.1 (Li et al. 2009) to filter alignments with a mapping quality < 20, and to sort and index the resulting alignments. To combine reads from one sample across multiple lanes, we used Picard tools 1.119 (http://broadinstitute.github.io/picard/) to edit BAM file headers and SAMtools 1.1 to merge, resort and index BAM files *per* sample. We then used Picard tools 1.119 to fix mate information, sort the resulting BAM files and mark duplicates. We performed local realignment using the RealignerTargetCreator and IndelRealigner tools of GATK 3.3 (https://www.broadinstitute.org/gatk/).

For SNP calling, we used the UnifiedGenotyper for diploid genomes (parameter: sample_ploidy 2) and generated a multisample VCF file (Danecek et al. 2011). Subsequently, we performed variant quality score recalibration (VQSR) to separate true variation from machine artefacts (DePristo et al. 2011). We used biallelic and homozygous (for a given individual) SNPs detected at 4-fold sites at a frequency equal to or higher than seven sequenced individuals as the training set. Six SNP call annotations were considered by the VQSR model: QD, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, and MQ, as suggested by GATK (see http://www.broadinstitute.org/gatk/; DePristo et al. 2011). The SNPs were allocated to tranches according to the recalibrated score, so that a given proportion of the true sites were recovered. We retained variants that passed a cutoff of 95%, that is, the variant score limit that recovers 95% of the variants in the true data set. We refer this dataset below as ‘filtered’. From the multisample recalibrated VCF file, we made a consensus sequence FASTA file for each line using a custom Perl script. The variant calls that did not pass the filter were called N (missing data) at the sites in question. We also generated an unfiltered dataset, where we did not implement any form of variant score recalibration. We refer to this dataset as ‘unfiltered’ below.

Annotation of the D. simulans *dataset*

Using annotations from the *D. simulans* reference (Hu et al. 2013) we extracted coding sequence (CDS) of each gene and made FASTA alignments. We included the *D.*

simulans reference sequence as well as the 1:1 FlyBase orthologous genes of *D. melanogaster* (release version 5.33) and *D. yakuba* (release version 1.3). We then performed amino-acid sequence alignments using MAFFT (Katoh et al. 2002). These amino-acid sequence alignments were translated back to nucleotides using custom scripts in PERL to produce in-frame coding sequence alignments that included the 42 *D. simulans* alleles, the *D. simulans* reference, and the *D. melanogaster* and the *D. yakuba* outgroups. We extracted 4-fold (and 0-fold) degenerate sites from CDS alignments which were 4-fold (0-fold) degenerate in all polymorphism sample lines with the condition that there was at most one segregating site in the codon to which the 4-fold (0-fold) site belonged. We retained the 4-fold (0-fold) sites from an alignment only if there were at least ten 4-fold (0-fold) sites in that alignment in total. For the polymorphism and substitution analyses on 4-fold sites reported below we carried out the same procedure with the added condition that sites must also be 4-fold degenerate in the three reference sequences.

We also extracted the intron coordinates from the *D. simulans* reference. Genomes were masked for any possible exons. For each *D. simulans* intron, we obtained the corresponding orthologous intron of *D. melanogaster* (Hu et al. 2013). For *D. yakuba*, for each orthologous gene, we obtained all its annotated introns and blasted them against the *D. melanogaster* introns (of the same ortholog) with an e-value of less than 1e-5 and selected the reciprocal best hit (because introns are generally short, the threshold e-value was conservative; see Results). We used RepeatMasker (http://www.repeatmasker.org) to mask repetitive elements in our intron dataset, using the library of repeats for *D. melanogaster* and the default settings. We produced a final alignment of each intronic polymorphic dataset of *D. simulans* with the corresponding *D. melanogaster* and *D. yakuba* orthologs using MAFFT.

We extracted positions 8-30bp of all introns < 66bp long, based on the *D. melanogaster* reference alignment for each intron, as we considered the *D. melanogaster* reference to be the best annotated of the three species. To do this, we scanned the *D. melanogaster* reference sequence for each intronic alignment. We retained the alignment if the *D. melanogaster* reference sequence was less than 66bp long (not including alignment gaps), and then further obtained the coordinates of the 8bp position and the 30bp position in the *D. melanogaster* reference sequence after discarding any gaps introduced by the alignment program. We then cut the whole alignment at these coordinates. These short intronic (henceforth 'SI') sites are thought to be the most neutrally evolving in

Drosophila based on patterns of polymorphism and substitution (Halligan and Keightley 2006; Parsch et al. 2010; Clemente and Vogl 2012b).

The D. melanogaster *dataset*

Similar analyses were performed using a *D. melanogaster* polymorphic dataset, which is described in in Jackson et al. (2015), consisting of 17 Rwandan *D. melanogaster* samples (RG18N, RG19, RG2, RG22, RG24, RG25, RG28, RG3, RG32N, RG33, RG34, RG36, RG38N, RG4N, RG5, RG7 and RG9) made available by the *Drosophila* Population Genomics Project (http://www.dpgp.org/dpgp2/candidate/).

Quality control of D. simulans *genotypes*

The lines that were inbred successfully for nine generations to produce homozygous samples still retained low levels of residual heterozygosity, which may have been due to a failure to purge our lines of natural variation (Stone 2012), or to SNP calling errors (the latter should be less likely given the high coverage [64x] and our stringent SNP calling regime). We quantified the amount of residual heterozygosity per sample for each of the unfiltered and filtered datasets (Supplementary Figure S1). As expected, the filtered dataset exhibited lower levels of residual heterozygosity (ND samples: mean value = 0.0616%, all values < 0.5%; MD samples: mean value = 0.0168%, all values < 0.15%). The six lines that were lost prior to inbreeding (see above) did not have substantially higher levels of residual heterozygosity than the remaining samples, potentially because they were already considerably inbred after being kept as laboratory stock for several years. For downstream analyses we treated heterozygous sites as follows: at each heterozygous site within a sample, one allele was chosen proportional to its coverage in that sample as the haploid genotype call at that site. The alternative allele was discarded. Because our samples are from partially inbred lines that originated from at least one wild male and only one wild female, heterozygosity at a site implies that site is segregating in the wild population. By sampling one allele, we intended to replicate the inbreeding process, by which we also aim to remove heterozygosity from within the lines.

Pairwise π_S values (synonymous site diversity) for all 42 *D. simulans* lines showed three pairs of samples which deviated substantially from the distribution of pairwise π_S between samples (mean π_S for all samples = 0.030, s.d. = 0.0018). These pairs were MD201—NS116 (π_S = 7.28e-5); NS137—NS37 (π_S = 0.0034) and NS49—NS96 (π_S = 0.0097). A PCA of binary genotypes placed NS116 within the cluster of MD samples, and NS116 exhibited more MD-like genetic distance to the *D. simulans* reference sequence. These results were based on the filtered dataset, but the unfiltered dataset returned qualitatively identical patterns (data not shown). We therefore excluded NS116 from all downstream analyses based on the likelihood of it representing labelling error. We also excluded NS37 and NS96 as these individuals had the highest levels of residual heterozygosity out of the remaining two pairs of closely related samples (Supplementary Figure S1).

To further assess the quality of our datasets, we compared polymorphism and divergence statistics to data previously published in the literature on *D. simulans* (see Results). In particular, we calculated a range of summary statistics per gene: F_{ST} between NS and MD samples; π, Tajima's D, Δ_π, and θ_W within the NS sample, within the MD sample, and for both samples combined. Δ_π (Langley et al. 2014) is defined as

$$\Delta_\pi = \frac{\hat{k}}{S} - \frac{1}{\sum_{i=1}^{n-1}(1/i)} \quad (1)$$

which we calculated using a modified version of the tajima.test() function from the pegas package (Paradis 2010) in R. Δ_π is similar to Tajima's D (Tajima 1989a), but is normalised by the total amount of diversity. Its advantage over Tajima's D is thus that it is less dependent on the total diversity of the sample (Langley et al. 2014). We also compared K_A and K_S between the three reference sequences (*D. melanogaster, D. simulans* and *D. yakuba*) in all CDS alignments using the kaks() function from the seqinr package in R, and K_{SI} between the reference sequences in all our short intronic alignments using the dist.dna() function from the pegas package in R, using the K80 method (Kimura 1980). These analyses are presented in the first section of the Results.

Divergence-based analyses

We used three methods to determine the ancestral state at the *melanogaster-simulans* (*ms*) node, all of which used only the three reference sequences. Firstly, we used parsimony, implemented in custom scripts in R. Secondly, we used the non-homogenous general time-reversible (GTR-NH_b) substitution model, implemented in the *baseml* package of PAML v4.8 (Yang 2007), after checking that GTR-NH_b fitted the data better than the stationary model GTR model using chi-squared tests (see Results). The use of this method to reconstruct ancestral sites when nucleotide composition is non-stationary is described in Matsumoto et al (2015), and has been shown to be able to produce highly accurate results in the presence of non-equilibrium base composition, in which case the parsimony method is likely to be biased. Under the GTR-NH_b method, we implemented two ways of determining the ancestral state at the *ms* node. Firstly, using the single best reconstruction (SBR) of the ancestral sequence at the *ms* node. Secondly, by weighting the four possible nucleotides at the *ms* node by the posterior probability of each. Instead of ignoring sub-optimal reconstructions, as the previous two methods do, this option weights all the possible ancestral states by their respective posterior probabilities. Following, Matsumoto et al. (2015), we refer to these two GTR-NH_b-based methods as 'SBR' and 'AWP' respectively. The AWP method should be more reliable than either parsimony or SBR when base composition is not at equilibrium (Matsumoto et al. 2015).

Since some of the models we used are very parameter-rich (e.g., the GTR-NH_b model has 39 parameters for three species, and the M1* model described below has 25 parameters for *simulans* and 21 parameters for *melanogaster*, given the sample sizes), we had to group genes into bins to avoid overfitting. To investigate the relationship between selection and GC content at 4-fold sites (a proxy for the extent of CUB), we binned 4-fold sites by the GC content in the *D. melanogaster* reference sequence, which we used as a proxy for the historic strength of selection favouring GC alleles. GC content evolves slowly over time (Marais et al. 2004), and is highly correlated between *D. simulans* and *D. melanogaster* CDS (data not shown), so this strategy should accurately represent GC content at the *ms* node. We binned 4-fold degenerate sites into 20 autosomal and 4 X-linked bins. Bins were chosen to maintain approximately the same number of genes per bin. The autosomal and X-linked SI sites were treated as two separate bins. We follow this binning convention for other analyses below. When carrying out correlation analysis between GC content bins and other variables (e.g.,

substitution rate and estimates of the selection coefficient [see below]), we included only the 4-fold degenerate site GC bins, but not the SI bin.

To determine whether or not *D. melanogaster* and *D. simulans* are in base composition equilibrium, we counted, for each bin, the number of $S \rightarrow W$ ($N_{S\rightarrow W}$), $W \rightarrow S$ ($N_{W\rightarrow S}$), and $neutral$ (N_{neu}) substitutions (i.e., $S \rightarrow S$ and $W \rightarrow W$), where S represents G or C, the strong (potentially preferred) allele, and W represents A or T, the weak (potentially unpreferred) allele. We will use GC and S interchangeably below; the same applies to AT and W. We did this along each of the *D. melanogaster* and *D. simulans* lineages using the reconstructed ancestral states at the *ms* node. For the AWP method, we rounded our results to the nearest integer. Where possible, we compared our results to those published in the literature, and to equivalent results kindly provided by Juraj Bergman and Claus Vogl (pers. comm.; supplementary Table S2). To obtain the $W \rightarrow S$ substitution rate ($r_{W\rightarrow S}$) per bin, we divided $N_{W\rightarrow S}$ by the total number of AT sites (L_W) at the *ms* node in that bin. Similarly, $r_{S\rightarrow W} = N_{S\rightarrow W}/L_S$.

Polymorphism-based analyses

For each bin, we estimated the derived allele frequency at segregating sites, using the three methods to infer ancestral states described above. We classified these sites into segregating sites at which the ancestral allele was AT and the derived allele was GC ($DAF_{W\rightarrow S}$), and segregating sites at which the ancestral allele was GC and the derived allele was AT ($DAF_{S\rightarrow W}$), as well as segregating sites which had mutated from A to T, or *vice versa*, and from G to C or *vice versa* (DAF_{Neu}). We also calculated Δ_π (Langley et al. 2014) per bin. We primarily display results obtained from the AWP method in the Results, because it is probably the most reliable of the three. Qualitatively, the results are generally insensitive to the choice of method for reconstructing ancestral sites. Thus, we present a set of figures in the supplement (Figures S5 – S11) that are parallel to those shown in the main text, but were obtained using either parsimony or SBR, respectively.

We used two methods for estimating the population-scaled strength of the force favouring GC alleles, $\gamma = 4N_e s$, where N_e is the effective population size and *s* is the selection coefficient against heterozygous carriers of the AT allele. Firstly, the method of Glémin et al (2015), which uses three different classes of polarised unfolded site frequency spectra (SFS) of sites which are segregating in the present day: $S \rightarrow W$, $W \rightarrow$

S, and $neutral$ (see above). This method is capable of taking into account polarization errors, which, if untreated, may lead to upwardly biases estimates of γ (Hernandez et al. 2007), by incorporating them into the model and estimating them jointly with the parameters of interest. It is also capable of accounting for demography, by introducing nuisance parameters to correct for distortions in the SFS due to demography (after Eyre-Walker et al. 2006). Because it only considers the SFS of derived alleles, we expect this method to recover signatures of selection on a relatively recent time scale ($\sim 4N_e$ generations, if we conservatively assume neutrality). We generated unfolded SFSes for this model using the AWP method to infer the ancestral state at the *ms* node, and estimated the strength of γ using R code provided in the supplement of Glémin et al (2015). We refer to the models using this method with the same notation as Glémin et al (2015). These are: model M0, where $\gamma = 0$ and polarisation errors are not taken into account; M1, where $\gamma \neq 0$ and polarisation errors are not taken into account; and M0* and M1*, which are the equivalent models accounting for polarisation errors. Note that the method for controlling for demography drastically increases the number model parameters. For instance, for M1, in addition to γ and the three mutational parameters for each of the three SFSes ($\theta = 4N_e\mu$), it requires an additional $n - 2$ nuisance parameters, where n is the number of frequency classes (in our case, this is the same as the sample size). Given the dearth of SNPs relative to substitutions, and in particular the lower diversity level in *D. melanogaster*, we repeated some of these analyses by pooling SNP data across several nearby GC content bins (see Results).

Secondly, we used the method of Zeng and Charlesworth (2009; see also Evans et al. 2014) which uses the unpolarised SFS (including fixed sites) to infer parameters of a two-allele model with reversible mutation between W and S alleles, selection and/or gBGC, and changes in population size (see Zeng (2012) for a discussion on the differences between the reversible mutation model and the infinite-sites model on which the method of Glémin et al. (2015) is based). Because this method uses the unpolarised SFS, no outgroup is required. This method may potentially recover signals of selection (and other population genetic parameters) over a longer time scale than the methods of Glémin et al (2015) (see Zeng and Charlesworth 2009 supplementary Figures S8 – S11). As above, we defined W (AT) and S (GC) as our two alleles. We define u as the rate at which S alleles mutate to W alleles, and v as the mutation rate in the opposite direction, and $\kappa = u/v$ as the mutation bias parameter. To incorporate a change in population size, we assume that the population in the past is at equilibrium with

population size N_1, which then changes instantly to N_0 (this can be either an increase or a reduction in size) and remains in this state for t generations until a sample is taken from the population in the present day (Zeng and Charlesworth 2009; Haddrill et al. 2011; Evans et al. 2014). As with M1* and M1, we also tested the equivalent models where $\gamma = 0$. For each model, in order to ensure that the true MLE was found, we ran the search algorithm multiple times (typically 500), each initialised from a random starting point. All the results reported below were found by multiple searches with different starting conditions. Chi-squared tests were used to evaluate statistical support for different models. We refer to these models as ZC0 ($\gamma = 0$) and ZC1 ($\gamma \neq 0$) below. A software package implementing this approach is available at http://zeng-lab.group.shef.ac.uk. For all methods (Zeng and Charlesworth 2009; and Glémin et al. 2015), we fitted independent models for each (SI and 4-fold) bin (Zeng and Charlesworth 2010b; Messer and Petrov 2013).

Results

Patterns of polymorphism and divergence in the D. simulans *and* D. melanogaster *datasets*

After extracting 4-fold degenerate sites and short introns, we retained 7551 autosomal CDS alignments and 1226 X-linked CDS alignments, as well as 5578 autosomal SI alignments and 516 X-linked SI alignments, containing polymorphism data from 39 *D. simulans* lines (21 MD lines and 18 NS lines) as well as outgroup data from the reference sequences of *D. simulans*, *D. melanogaster* and *D. yakuba*. For *D. melanogaster*, we retained 5550 autosomal CDS alignments and 888 X-linked autosomal alignments, as well as 7397 autosomal SI alignments and 738 X-linked SI alignments, containing polymorphism data from 17 RG lines, as well as the three reference sequences.

Summary statistics calculated using the filtered *D. simulans* data are presented in Table 1 (see supplementary Table S1 for the unfiltered data). Consider first the MD lines (n = 21) collected from the putatively ancestral range of the species (Dean and Ballard 2004). Autosomal π at 4-fold sites (referred to as π_4) was 0.0329 and 0.0317 for the unfiltered and filtered datasets, respectively, similar to the value of 0.035 reported by Begun et al (2007). On the X, π_4 was 0.0191 and 0.0182 for the two datasets, and Begun et al's (2007) was 0.02. Tajima's D and Δ_π at 4-fold sites are both negative, implying

that there may have been a substantial recent population size expansion. Again, values obtained from the filtered and unfiltered data are very similar (cf. Tables 1 and S1). Overall, diversity was slightly reduced for our filtered dataset, which may have been a result of more conservative masking criteria, but the differences are minimal. In what follows, we only present results obtained from the filtered dataset. SI sites, which we only obtained from our filtered dataset, are more diverse than 0-fold and 4-fold sites in the MD population, and on both the A (π_{SI} = 0.0321) and the X (π_{SI} = 0.0208) (Table 1).

Samples collected from Kenya (the NS lines; n = 18) have consistently lower diversity level at 0-fold, 4-fold and SI sites, and less negative Tajima's D and Δ_π, which may be caused by bottlenecks associated with the colonisation process (Dean and Ballard 2004). Nonetheless, F_{ST} at 4-fold sites is rather low: ~2.5% between NS and MD (Table 1), suggesting that there is relatively little genetic differentiation between the ancestral and derived populations. There is also little difference in F_{ST} at 4-fold sites between the X and autosomes. As with the MD population, SI sites are the most diverse class of site according to π (Table 1).

The patterns reported above contrast with those observed in *D. melanogaster* (see Table 1 of Jackson et al. 2015). Focusing first on samples from the putatively ancestral ranges of both species (i.e., RG vs MD), autosomal π_4 is ~2.06 times higher in *D. simulans*, suggestive of higher N_e, which may in turn lead to more efficient selection (see Discussion). Tajima's D is also less negative in *D. melanogaster*, with the differences at 4-fold sites being the most noticeable (-0.11 *vs* -1.03 for the autosomes, and -0.47 *vs* -1.31 for the X), suggesting a more stable recent population size in *D. melanogaster*. The X:A ratio of π_4 in *D. melanogaster* was 1.08, much higher than the expected value of 0.75 under the standard neutral model, whereas it was 0.57 in *D. simulans*. Furthermore, F_{ST} at 4-fold sites between RG and a sample from France (see Jackson et al. 2015 for details) in *D. melanogaster* is ~10 times higher than that between the MD and NS populations in *D. simulans.* Interestingly, the difference in F_{ST} between the X and autosomes is much more marked in *D. melanogaster* (0.29 vs. 0.17 for the X and autosomes, respectively) than in *D. simulans* (0.025 for both X and A). Various theories have been proposed to explain differences between X and autosomes, which include sex-specific variance in reproductive success (Charlesworth 2001), demographic effects (Pool and Nielsen 2007; Singh et al. 2007; Pool and Nielsen 2008; Yukilevich et al. 2010), positive and negative selection (Singh et al. 2007; Charlesworth 2012a), and differences in recombination rate (Charlesworth 2012a). Detailed analyses of the factors

underlying X-autosomal differences are outside the scope of this study; below we present results from X and the autosomes separately.

We also assayed divergence between the reference sequences in our alignments. Between *D. melanogaster* and *D. simulans*, K_A, K_S and K_{SI} were 0.014, 0.109 and 0.130, respectively. These values were similar to those in Table 1 of Parsch et al (2010) (K_A = 0.019, K_S = 0.106 and K_{SI} = 0.123), and those in Zhang et al (2013; Supplementary Table 2) (K_A = 0.015 and K_S = 0.12). In our data K_A, K_S and K_{SI}, between *D. melanogaster* and *D. yakuba* were 0.036, 0.266 and 0.294, respectively; between *D. simulans* and *D. yakuba*, they were 0.036, 0.250 and 0.302, respectively. Note that divergence is always highest at the SI class of site, which is in agreement with these sites being relatively unconstrained (Halligan and Keightley 2006; Parsch et al. 2010; Clemente and Vogl 2012b). Overall, these data suggest that our alignments are of high quality.

A thorough investigation of the causes of these contrasting patterns between the two *Drosophila* species is beyond the scope of this paper. In the remaining sections, we will focus on studying forces that act on 4-fold sites. The putatively neutrally-evolving SI sites are analysed separately and presented alongside results from 4-fold sites for comparison. Only data from the ancestral populations (i.e., MD in *D. simulans* and RG in *D. melanogaster*) are considered to avoid complications induced by population structure.

Excess of $S \rightarrow W$ substitutions in both the D. simulans *and the* D. melanogaster *lineages*

For all autosomal, X-linked, CDS and SI bins, the non-homogenous GTR-NH_b substitution model implemented in PAML always fitted the data significantly better than the stationary GTR substitution model (min χ^2 = 166.86, d.f. = 28, p = 1.05e-21), indicative of non-equilibrium base composition evolution. Both the *D. melanogaster* and *D. simulans* lineages showed an excess of $S \rightarrow W$ changes ($N_{S \rightarrow W}$) at autosomal and X-linked 4-fold degenerate sites, regardless which method was employed to infer ancestral states at the *ms* node (Tables 2 and S2). It is evident that the excess is greater in *D. melanogaster* than *D. simulans*. For instance, based on autosomal data obtained by the AWP method, $N_{W \rightarrow S}/N_{S \rightarrow W}$ is 0.49 in *D. simulans*, but 0.26 in *D.* melanogaster, which is significantly lower (χ^2 = 2145.8, d.f. = 1, $p < 0.001$). The $S \rightarrow W$ bias is much more pronounced on the X of *D. melanogaster* with an $N_{W \rightarrow S}/N_{S \rightarrow W}$ ratio of 0.17,

significantly different from 0.26 on the autosomes (χ^2 = 212.8, d.f. = 1, p < 0.001), whereas in *D. simulans* the ratios are much closer to one another, 0.53 and 0.49, respectively, although this difference is still significant (χ^2 = 6.97, d.f. = 1, p = 0.008). These results are in line with previous findings of an excess of AT (or unpreferred codon) substitutions at silent sites in *D. melanogaster* (Akashi 1995; Akashi 1996; Takano-Shimizu 2001; Akashi et al. 2006). For *D. simulans*, our data are in agreement with a dataset curated entirely independently by Juraj Bergman and Claus Vogl (pers. comm.; supplementary Table S2), and suggest that there is a much more pronounced $S \rightarrow W$ bias than was found in some previous studies (Akashi et al. 2006; Begun et al. 2007; Poh et al. 2012).

The ratio of $N_{W\rightarrow S}/N_{S\rightarrow W}$ is much closer to unity for SI sites than for 4-fold sites (Tables 2 and 3), which is also in agreement with previous results that found short introns are generally closest to equilibrium in both species (Kern and Begun 2005; Haddrill and Charlesworth 2008; Singh et al. 2009; Robinson et al. 2014). The three methods for inferring ancestral states in the *ms* ancestor consistently suggested an AT substitution bias at SI sites in the *D. melanogaster* lineage (Table 3). The situation is somewhat more complex in *D. simulans.* For the X, all three methods suggest a mild GC bias, but the ratio based on AWP, which should be the most reliable method of the three (Matsumoto et al. 2015), is not significantly different from 1 (χ^2 = 0.237, d.f. = 1, p = 0.63). For the autosomes, parsimony suggests a GC bias (χ^2 = 19.7, d.f. = 1, p = 0.01), but both SBR and AWP provide some support for a slight AT bias (SBR: χ^2 = 3.73, d.f. = 1, p = 0.05; AWP: χ^2 = 5.55, d.f. = 1, p = 0.019) (Table 3). This may reflect the tendency for parsimony to overestimate common to rare changes (Collins et al. 1994; Eyre-Walker 1998; Akashi et al. 2007; Matsumoto et al. 2015).

Variation in substitution patterns across regions with different GC content

Under strict neutrality, the substitution rate per site is equal to the mutation rate per site (Kimura 1983). Thus, if 4-fold degenerate sites have never been affected by selection on CUB and/or gBGC, the two substitution rates per site, $r_{W\rightarrow S}$ and $r_{S\rightarrow W}$, should be uniform across the GC bins. However, as can be seen from Figure 1, in both species, on both the autosomes and the X chromosome, $r_{W\rightarrow S}$ is significantly positively correlated with GC content (*D. simulans*, autosomes: Kendall's τ = 0.45, p = 0.006; *D. melanogaster*, autosomes: τ = 0.53, p = 0.001; here and in what follows, we refrain from

conducting formal correlation tests of the X-linked data due to the dearth of data points; in all cases, data from the SI bins were not included in correlations), whereas $r_{S\to W}$ shows a clear negative relationship with GC content (Kendall's τ = -0.95, $p < 0.001$ and τ = -0.96, $p < 0.001$ in *D. simulans* and *D. melanogaster* autosomes, respectively). These patterns are expected if GC alleles (i.e., preferred codons) were favoured over AT alleles (i.e., unpreferred codons) for a substantial amount of time along these two lineages, and that the intensity of the GC-favouring force increases with GC content (see Discussion for an explicit model). Also of note is the marked increase in $r_{S\to W}$ relative to $r_{W\to S}$ in the *melanogaster* lineage, which is suggestive of mutation becoming more AT-biased. However, the arguments set out in Discussion suggest that a change in mutational bias alone is unlikely to explain the data reported here.

As stated before, the $N_{W\to S}/N_{S\to W}$ ratio at SI sites, particularly in *D. simulans*, is close to unity, the value expected under equilibrium base composition. An investigation across the 4-fold site GC content bins suggests that all of the bins considered here are experiencing some level of AT fixation bias ($N_{W\to S}/N_{S\to W} < 1$), and that genomic regions with higher GC contents are evolving towards AT faster than regions with lower GC contents. This is clear from the negative correlations between GC content and the level of substitution bias ($N_{W\to S}/N_{S\to W}$) calculated per 4-fold site bin in both species (Kendall's τ = -0.96, $p < 0.001$ and τ = -0.91, $p < 0.001$ in *D. simulans* and *D. melanogaster* autosomes, respectively) (Figure 2). As explained in the Discussion, this negative correlation can most easily be explained by a genome-wide reduction in the intensity of the GC-favouring force.

Derived allele frequencies (DAF) provides clear evidence of ongoing selection for preferred codons

If selection/gBGC favours GC alleles over AT alleles, then frequencies of derived GC alleles at AT/GC polymorphic sites ($DAF_{W\to S}$) should on average be higher than frequencies of derived AT alleles at AT/GC polymorphic sites ($DAF_{S\to W}$). Furthermore, $DAF_{W\to S}$ should increase as the GC-favouring force becomes stronger (i.e. as 4-fold site GC content increases), whereas $DAF_{S\to W}$ should decrease with increasing GC content. In addition, we expect the DAF for $neutral$ changes (DAF_{neu}) to lie in a position intermediate between $DAF_{S\to W}$ and $DAF_{W\to S}$ (i.e., $DAF_{W\to S} > DAF_{neu} > DAF_{S\to W}$). In

contrast, in a neutral model with a recent increase in mutational bias towards AT, the higher number of derived AT mutations entering the population, which tend to be young and segregate at low frequencies, will depress $DAF_{S\to W}$, leading to $DAF_{W\to S} > DAF_{S\to W}$, but DAF_{neu} should be comparable to $DAF_{W\to S}$. Moreover, GC content and $DAF_{W\to S}$ should be unrelated under this model.

D. simulans fits the expectations of the first model: $DAF_{W\to S}$ was greater than $DAF_{S\to W}$ in all autosomal and X-linked 4-fold bins, and DAF_{neu} was always intermediate between $DAF_{W\to S}$ and $DAF_{S\to W}$ (Figure 3). Autosomal 4-fold site $DAF_{W\to S}$ correlated positively with GC content (Kendall's $\tau = 0.6$, $p < 0.001$; Figure 3), and autosomal 4-fold site $DAF_{S\to W}$ correlated negatively with GC content (Kendall's $\tau = -0.85$, $p < 0.001$; Figure 3); data from the X displayed similar trends. These patterns suggest the action of forces favouring GC over AT alleles in the recent past in this species (of the order of $4N_e$ generations), with higher GC content bins experiencing a higher strength of recent selection favouring GC.

In *D. melanogaster*, the equivalent results are less clear. Autosomal $DAF_{W\to S}$ was higher than autosomal $DAF_{S\to W}$ for 19/20 4-fold bins (Figure 3). As in *D. simulans*, autosomal 4-fold $DAF_{W\to S}$ correlated positively with GC content (Kendall's $\tau = 0.41$, $p = 0.01$; Figure 3), and autosomal 4-fold $DAF_{S\to W}$ correlated negatively with GC content (Kendall's $\tau = -0.47$, $p = 0.004$; Figure 3). DAF_{neu} fell between $DAF_{W\to S}$ and $DAF_{S\to W}$ in 14/20 autosomal 4-fold site bins, but only 1/4 X-linked 4-fold bins (Figure 3). Additionally, the difference between $DAF_{W\to S}$ and $DAF_{S\to W}$ seems less pronounced than in *D. simulans*, especially on the X chromosome, although on the autosomes, the gap between $DAF_{W\to S}$ and $DAF_{S\to W}$ does tend to increase with GC content and is largest and comparable in magnitude to those seen in *D. simulans* in the bins with highest GC content. Overall, these data provide some evidence of recent selection for GC at 4-fold sites in *D. melanogaster*, but its extent seems to be less than in *D. simulans*, and may be restricted to autosomal regions with high GC contents.

Estimating γ using polymorphism data

To shed further light on the evolutionary dynamics of selection on CUB, we used two different methods to infer γ from polymorphism data, with the Glémin method detecting recent selection ($\sim 4N_e$ generations), and ZC providing estimates over a longer period.

For every *D. simulans* bin on both the autosomes and X, both ZC1 and M1 fitted the data significantly better than the corresponding models with $\gamma = 0$ (i.e., ZC0 and M0; min $\chi^2 = 17.84$, d.f. = 1, $p < 0.001$); the only exception is the X-linked SI bin where M1 does not fit the data better than M0 ($\chi^2 = 0.071$, d.f. = 1, $p = 0.79$) (Figure 4). Estimates obtained by ZC1 and M1 agreed closely in their results for the *D. simulans* data (Figure 4). The alignment of the results from the two methods, which are expected to be sensitive to forces favouring GC on different timescales, suggests consistent selection over time favouring GC alleles at 4-fold degenerate and SI sites in *D. simulans*. In addition, GC content correlated positively with γ on both the autosomes (Kendall's $\tau = 0.98$, $p < 0.001$; $\tau = 0.88$, $p < 0.001$ for ZC1 and M1, respectively) and the X chromosome. Thus, in agreement with the results obtained from the divergence- and DAF-based analyses, selection for GC is indeed stronger in regions with higher GC content. The patterns obtained from comparing M0* and M1* were qualitatively identical (Supplementary Figure S2). In addition, when using the Akaike Information Criterion (AIC) to rank the four models (this is necessary because, e.g., M0* and M1 are not nested and cannot be compared using the likelihood ratio test), M1 and M1* are always the two best fitting models for all bins across both chromosome sets, except for the SI bin on the X (Supplementary Table S3).

Similarly to the analysis based on DAFs, the patterns are less clear-cut in *D. melanogaster*. When M1 and M0 were compared, 13/20 autosomal 4-fold site bins were found to be non-neutrally evolving, including the four highest autosomal GC bins, and none on the X (Figure 4). In contrast, according to the comparison between M1* and M0*, only 3 autosomal bins showed evidence of non-zero γ in *D. melanogaster* (2/20 autosomal 4-fold site bins and the autosomal SI bin), and none of the X-linked bins did so (Supplementary Figure S2). In particular, the fact that none of the high GC bins have a significant test is out of keeping with the observation that these bins have large differences between $DAF_{W \to S}$ and $DAF_{S \to W}$. A close inspection suggests that statistical power may be an issue: there are on average 4-times fewer SNPs in the 4-fold site bins in *D. melanogaster*, and in the highest 4-fold site bin, there were only 69 $W \to S$ SNPs. As described in Methods, the Glémin models are parameter-rich, especially M0* and M1*. In fact, M1* often came out (e.g., in 10/20 autosomal 4-fold site bins) as the worse fitting one among the four models according to the AIC. To deal with this issue, we redid the comparison by reducing the number of autosomal 4-fold bins to 10. M1 fits better than M0 in 9/10 bins, while M1* fits better than M0* in 4/10 bins, including 2

out of the top 4 GC bins (Supplementary Figure S3). According to the AIC, the frequency of M1 being the best fitting model increased to 9/10 bins, whereas the frequency of M1* being the worse fitting model decreased to 2/10 bins (Supplementary Table S3). The observation that M1* sometimes ranked lower than M1 according to the AIC in both species may also be due to the fact that our method to account for non-equilibrium when reconstructing ancestral states has reduced the need to account for polarisation errors.

As is apparent from Figure 4, M1 also estimated consistently lower absolute values of γ than ZC1 in *D. melanogaster*. Given that the ZC method returns long-term average estimates of γ, these differences clearly indicate recent declines in the strength of selection on CUB in this species. As with *D. simulans*, however, autosomal GC content was correlated positively with γ under both models (Kendall's $\tau = 0.87$, $p < 0.001$; $\tau = 0.48$, $p = 0.003$ for ZC and M1, respectively; Figure 4), which is suggestive of some, if weak, ongoing selection for GC at autosomal 4-fold sites, particularly in GC-rich regions of the genome. The fact that the SFS is more negatively skewed, as measured by Δ_π at 4-fold sites in regions of higher GC content in both species (Supplementary Figure S4) also supports the idea that selection is acting at these sites.

Discussion

Evidence of past selection on CUB in both Drosophila species

The correlations between the substitution rates and GC content presented in Figures 1 and S5 can be explored by using the following modelling framework (Li 1987; Bulmer 1991; McVean and Charlesworth 1999), which assumes a fixed N_e and thus a fixed value of γ for each GC bin. If there are temporal changes along a lineage, we can regard these parameters as long-term averages. Let u be the mutation rate from $S \rightarrow W$ per site per generation; and v be that in the opposite direction. Define κ as u/v. It can be shown that the two substitution rates, $r_{S \rightarrow W}$ and $r_{W \rightarrow S}$, are proportional to $u\gamma/[exp(\gamma) - 1]$ and $v\gamma/[1 - exp(-\gamma)]$, respectively (e.g., Eq. B6.4.2b of Charlesworth and Charlesworth (2010); Eq. 11 of Sawyer and Hartl (1992)). We can then define

$$R = \frac{r_{S\to W}}{r_{W\to S}} = \kappa \frac{1 - e^{-\gamma}}{e^{\gamma} - 1} = \kappa e^{-\gamma} \quad (2)$$

Assuming that u and v are constant across the GC bins and over time (κ is thus also constant), R is a function of γ. Taking the derivative with respect to γ, we have

$$\frac{dR}{d\gamma} = -\kappa e^{-\gamma} \quad (3)$$

In other words, $R = \kappa$ when $\gamma = 0$ (neutrality), and decreases as γ becomes positive (i.e., when W is selected against). Thus, the decreasing values of R shown in Figures 1 and S5 suggest that S is more strongly favoured in high GC bins. For instance, R calculated for the lowest (bin 1) and highest (bin 20) autosomal 4-fold site bins in *D. simulans* is 1.51 and 0.56, respectively. Assuming that κ is about 2 (Singh et al. 2005; Keightley et al. 2009; Zeng 2010; Schrider et al. 2013), solving Eq. (2) for γ gives 0.28 and 1.27, respectively.

In contrast, if we assume that $\gamma = 0$ across the bins (i.e., assuming that there has been no selection along both the *D. melanogaster* and *D. simulans* lineages), since $R = \kappa$, a genome-wide increase in κ (i.e., a more AT-biased mutation pattern) will not cause a negative relationship between R and GC content. In addition, the trends of $r_{S\to W}$ and $r_{W\to S}$ shown in Figure 1 suggest that, if the relationship between R and GC content were entirely due to variation in the mutation rates, then u decreases as GC content increases, whereas v changes in the opposite direction. However, there is little evidence to suggest that the mutation rate varies in this manner. Additionally, such a neutral model with variation in mutational bias is incompatible with evidence of selection from the two polymorphism-based methods (Figure 4), and cannot easily explain the well-known positive correlation between GC content of coding sequences (or the extent of CUB) and gene expression levels (e.g., Campos et al. 2013), especially when considering the lack of support for transcription-coupled repair in *Drosophila* (Singh et al. 2005; Keightley et al. 2009). Put together, it seems that the patterns regarding R shown in Figure 1 can be most parsimoniously explained by the existence of forces favouring GC for a substantial fraction of the time along the *melanogaster* and *simulans* lineages. As detailed in the supplementary text, this model can also explain why the slope for $r_{S\to W}$ is apparently steeper than that $r_{W\to S}$ (Figure 1).

The above model can also explain why, at 4-fold sites, $R_N = N_{W\to S}/N_{S\to W} < 1$ and there is a negative relationship between R_N and GC content (Figure 2), where $N_{W\to S}$ and $N_{S\to W}$ are the numbers of substitutions between the S and W alleles along the lineage of interest, respectively. Note first that $N_{S\to W}$ and $N_{W\to S}$ are, respectively, proportional to $Qu\gamma/[exp(\gamma)-1]$ and $(1-Q)v\gamma/[1-exp(-\gamma)]$, where Q is the GC content at the *ms* node (since Q changes very slowly, this should be a reasonable first approximation). At equilibrium, $Q = 1/[1+\kappa exp(-\gamma)]$ (Li 1987; Bulmer 1991) and hence $N_{W\to S}/N_{S\to W} = 1$. Consider a model where the ancestral species is at equilibrium, but γ reduces to $p\gamma(0 \le p < 1)$ along a lineage that leads to an extant species, so that $N_{S\to W}$ and $N_{W\to S}$ become proportional to $Qup\gamma/[exp(p\gamma)-1]$ and $(1-Q)vp\gamma/[1-exp(-p\gamma)]$. Then, R_N for the GC content bin in question can be written as

$$R_N = \frac{N_{W\to S}}{N_{S\to W}} = \frac{(1-Q)(e^{p\gamma}-1)}{\kappa Q(1-e^{-p\gamma})} = e^{-(1-p)\gamma} \quad (4)$$

Assuming that p is constant across bins (i.e., there has been a genome-wide proportional reduction in γ), then R_N decreases as γ increases. Thus, together with the arguments presented above that the long-term average γ is higher in high GC bins, Eq. (4) suggests that an explanation of the negative relationship between R_N and GC content is a genome-wide reduction in the intensity of selection.

Overall, the above theoretical explorations suggest that the data presented in Figures 1, 2 and S5 cannot be explained by a shift towards a more AT-biased mutational pattern alone; instead selection preferring GC over AT must have acted on both species for a significant amount of time since they last shared a common ancestor. However, it should be noted that this does not exclude the possibility that mutation has become more AT-biased. Indeed, comparing the autosomal data from *D. simulans* and *D. melanogaster* in Figure 1, while values of $r_{W\to S}$ are broadly similar between *D. simulans* and *D. melanogaster*, there is an apparent increase in $r_{S\to W}$ in the *melanogaster* lineage, which may have been caused by more frequent mutations towards AT, as has been suggested by previous authors (Takano-Shimizu 2001; Kern and Begun 2005; Nielsen et al. 2007; Zeng and Charlesworth 2010a; Clemente and Vogl 2012b).

Estimating the intensity of selection on preferred codons on different timescales

A novelty of this study is that we have attempted to understand how the selective pressure on codon usage has changed over time by using methods that can estimate γ on either a short timescale; (for roughly the last $4N_e$ generations; i.e., the Glémin method [Glemin et al. 2015]) , or a long timescale (i.e., the ZC method [Zeng and Charlesworth 2009]). The time period considered by the ZC method should be at least $\sim 4N_e$ generations. However, pinpointing the exact timescale is difficult, because it depends on details of past evolutionary dynamics that we know little about (e.g., the timescale can be affected by both when the ancestral population size reduction took place and the severity of the reduction; see supplementary Figures S8 – S11 in Zeng and Charlesworth 2009). This difference in timescale between the methods is due to the use of the derived SFS under the infinite-sites model in the Glémin method (Kimura 1983; Charlesworth and Charlesworth 2010) and the use of a reversible mutation model in the ZC method (see Zeng 2012 for a more thorough discussion on the differences of these two models and their application to the study of CUB). By the same token, we can classify other polymorphism-based methods into short timescale (Akashi and Schaeffer 1997; Bustamante et al. 2001) and long timescale (Maside et al. 2004; Cutter and Charlesworth 2006; Galtier et al. 2006; Zeng 2010; Clemente and Vogl 2012a; Vogl and Bergman 2015).

Contrasting the results obtained from the ZC method with those from the divergence-based analysis (Figures 1 and 2) and the Glémin method (Figure 4) is informative. Consider firstly *D. simulans*. The fact that values of γ estimated by both the ZC method and the Glémin method are virtually identical suggests that there have not been significant changes in the intensity of selection over the time period that the ZC method considers. Hence, the reduction in γ suggested in the previous section, which may have caused $N_{W\to S}/N_{S\to W} < 1$ and the negative correlation between $N_{W\to S}/N_{S\to W}$ and GC content, should have happened sufficiently early during the evolution of *D. simulans* that it did not leave detectable traces in the polymorphism data. In contrast, in *D. melanogaster*, both the divergence-based analysis and the comparison between the ZC method and the Glémin method provide evidence of reduction in γ, indicating a sustained decline. Assuming that short introns are neutral, and using autosomal data from the putatively ancestral populations (i.e., MD and RG), Table 1 in this study and Table 1 in Jackson et al. (2015) suggest that N_e is 2.21-fold higher in *D. simulans*, implying more efficient selection. In fact, focusing on the 13 autosomal 4-fold site bins

in *D. melanogaster* where M1 fits the data better than M0 (filled squares in Figure 4), the γ estimates in the corresponding bins in *D. simulans* is on average 2.93 times higher, comparable to the difference in N_e suggested by the short intron data. This difference in N_e may be due to differences in the two species' demographic history (i.e., frequencies of population bottlenecks). Previous studies have also suggested that the lower recombination rate in *D. melanogaster* compared to *D. simulans* (True et al. 1996; Comeron et al. 2012) may have played a role through stronger Hill-Robertson interference between selected sites (Takano-Shimizu 1999; McVean and Charlesworth 2000; Comeron et al. 2008; Comeron et al. 2012; Cutter and Payseur 2013). However, without detailed genetic maps from closely-related outgroup species, it is impossible to ascertain whether the reduced map length in *D. melanogaster* represents the ancestral or derived state; this is an important area for further research.

Potential effects of other evolutionary forces

In the above discussion, we have assumed the GC is favoured over AT because of selection for preferred codons. As mentioned in the Introduction, a factor that could confound the study of CUB is GC-biased gene conversion (Duret and Galtier 2009). Generally, studies have found little or no evidence for gBGC in *D. melanogaster* (Clemente and Vogl 2012b; Campos et al. 2013; Robinson et al. 2014). In particular, on examining a detailed genetic map, Comeron et al. (2012) failed to find a positive correlation between gene conversion rate and GC content, as predicted by the gBGC model. It is therefore somewhat surprising to see that the data presented in Figures 3 and 4 suggest that GC is favoured over AT in short introns (see also supplementary Figures S10c and S11c). However, in a recent analysis of non-coding regions in *D. melanogaster*, Robinson et al. (2014) also detected some sporadic support for increased fixation probability of $W \rightarrow S$ mutations in short introns, although the overall pattern is incompatible with the gBGC model (e.g., there is no evidence that GC is more favoured in regions with higher recombination rates). It is possible that these sporadic signals, when combined into a single SI bin, have contributed to the results reported here. More investigations are needed to identify what may have led to these patterns in short introns.

Less is known about *D. simulans* with respect to gBGC, although an earlier analysis of X-linked non-coding sequences (Haddrill and Charlesworth 2008) obtained patterns that

are analogous to the short intron data shown for *D. simulans* in Figure 3, suggestive of GC being favoured over AT in these regions. However, the current lack of a high-resolution genetic map with refined estimates of the rates of crossover and gene conversion (Comeron et al. 2012) hinders progress towards a deeper understanding of the relative importance of gBGC in *D. simulans*. For instance, we do not know whether gene conversion rate is positively related to GC content. Nonetheless, gBGC is unlikely to be the sole explanation of our results. From Figure 4, γ estimates obtained from 4-fold sites are consistently higher than the short intron estimate and increases to more than 4-fold higher in high GC regions. More generally, the gBGC model cannot easily explain the observation that gene expression level is strongly positively correlated with the GC content of coding sequences, but shows no relationship with intronic GC content, reported in *D. melanogaster* (Campos et al. 2013). This issue should be investigated in *D. simulans* in the future when high-quality expression data become available.

A final factor worth considering is the suggestion that a subset of 4-fold sites may be under strong selective constraints in *D. melanogaster* (Lawrie et al. 2013). These authors based their conclusions on two main observations that were made from analysing a North American population generated by the *Drosophila* Genetic Reference Panel (DGRP): a lack of difference in the shape of the SFSs between 4-fold and SI sites and a ~22% reduction in diversity level at 4-fold sites relative to SI sites (after correcting for differences in GC content; see their Figure 1). The authors suggested that their findings might represent "a largely orthogonal force to canonical codon usage bias" (p. 12 in Lawrie et al. 2013). Indeed, by using a sample of size 130, they were able to detect signals of much stronger purifying selection (with γ estimated to be -283) than permitted by our sample sizes (21 MD lines from *D. simulans* and 17 RG lines from *D. melanogaster*). Additionally, their estimates of the intensity of strong selection appear to be fairly uniform across genes with high and low levels of CUB, in contrast to the pattern we report here.

Obtaining more information about these two seemingly independent forces acting on 4-fold sites is an important area for future investigation. Several factors are of note. The North American population of *D. melanogaster* is known to have formed by admixture between African and European flies (Caracristi and Schlötterer 2003; Bergland et al. 2016). Although Lawrie et al. (2013) used the same method as Glémin et al. (2015) to control for demography, this method is nonetheless an approximation and may still lead

to biased estimates of γ under certain conditions, as demonstrated by simulations (Eyre-Walker et al. 2006). Using non-admixed populations (as in this study) and explicit demographic models (as in the ZC method) may be preferable. Second, with a larger sample size (as in Lawrie et al. 2013), it should be possible to jointly model the effects of both weak selection on CUB, which requires distinguishing $W \rightarrow S$, $S \rightarrow W$, and neutral mutations (i.e., $S \rightarrow S$ and $W \rightarrow W$) (these were ignored by Lawrie et al. 2013), and strong purifying selection, which primarily leads to an excess of very low-frequency variants. By doing so, we should be able to explicitly test the relative importance of these two forces, and gain further insights into the evolution of 4-fold sites in the *Drosophila* genome.

Tables

Table 1. Summary statistics for the filtered *D. simulans* dataset

Chr[a]	Site	Within population					Between populations	Total no. sites
		Pop.[b]	π[c]	θ_W[d]	Δ_π[e]	D[f]	F_{ST}	
A	0-fold[g]	MD	0.0016	0.0027	-0.120	-1.290	0.0202	6,841,064
		NS	0.0015	0.0021	-0.088	-0.903		
	4-fold[h]	MD	0.0317	0.0434	-0.078	-1.030	0.0252	1,555,042
		NS	0.0294	0.0347	-0.046	-0.579		
	SI[i]	MD	0.0321	0.0417	-0.065	-0.603	0.0174	110,758
		NS	0.0297	0.0340	-0.036	-0.326		
X	0-fold	MD	0.0012	0.0021	-0.125	-1.270	0.0178	1,095,517
		NS	0.0011	0.0016	-0.094	-0.924		
	4-fold	MD	0.0182	0.0282	-0.104	-1.310	0.0246	261,476
		NS	0.0173	0.0225	-0.071	-0.847		
	SI	MD	0.0208	0.0298	-0.092	-0.785	0.0194	10,884
		NS	0.0195	0.0248	-0.059	-0.509		

All statistics were calculated per gene, and the means are presented here.

[a] Chromosome

[b] Population sample: MD – Madagascar; NS – Kenya

[c] Average number of pairwise differences between lines

[d] Watterson's estimator of θ, the scaled mutation rate

[e] See equation (1)

[f] Tajima's D

[g] 0-fold degenerate sites

[h] 4-fold degenerate sites

[i] Sites 8-30bp of introns < 66bp in length

Table 2. Counts of substitutions along the *D. melanogaster* and *D. simulans* lineages at 4-fold degenerate sites. The ancestral state at the *melanogaster-simulans* node was determined using three methods: parsimony, the single best reconstruction (SBR) under the GTR-NH_b model implemented in PAML, and the average weighted by posterior probability (AWP) under the GTR-NH_b model implemented in PAML.

	D. simulans				*D. melanogaster*			
	A		X		A		X	
Polarisation method	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC	GC → AT
parsimony	13607	25656	1962	3934	10588	40586	1140	7395
SBR	14085	30524	2116	4528	11285	47894	1258	8670
AWP	15219	30945	2450	4639	12399	48264	1425	8611

Table 3. Counts of substitutions along the *D. melanogaster* and *D. simulans* lineages at short intronic (SI) sites. The ancestral state at the *melanogaster-simulans* node was determined using three methods: parsimony, the single best reconstruction (SBR) under the GTR-NH_b model implemented in PAML, and the average weighted by posterior probability (AWP) under the GTR-NH_b model implemented in PAML.

	D. simulans				*D. melanogaster*			
	A		X		A		X	
Polarisation method	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC	GC → AT
parsimony	1859	1598	206	152	1570	1884	131	229
SBR	1930	2052	231	183	1658	2417	146	271
AWP	2006	2158	217	206	1718	2506	141	303

Figures

Figure 1. Substitution rates for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant *D. melanogaster* reference sequence. Rates were calculated for the *D. simulans* lineage (top row) and the *D. melanogaster* lineage (bottom row), for autosomes (left-hand column) and X-linked sites (right-hand column). AT → GC substitutions – teal circles; GC → AT substitutions – orange triangles.

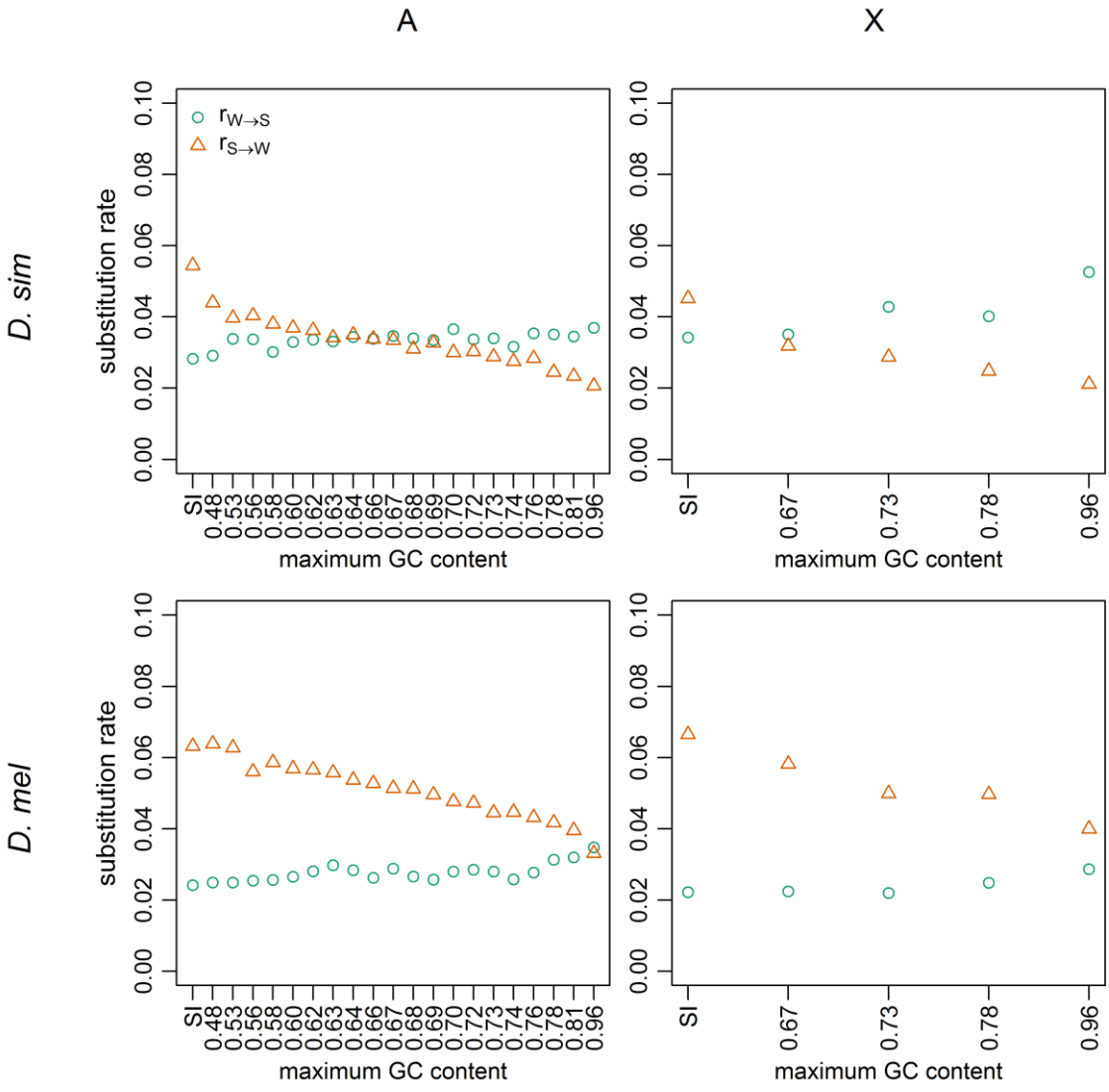


Figure 2. The ratio of substitution counts for positions 8-30bp of introns <66bp long (SI sites; leftmost point), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant *D. melanogaster* reference sequence. A substitution count ratio of $N_{W\to S}/N_{S\to W} = 1$ implies equilibrium base composition. Ratios were calculated for the *D. simulans* lineage (top row) and the *D. melanogaster* lineage (bottom row), for autosomes (left-hand column) and X (right-hand column).

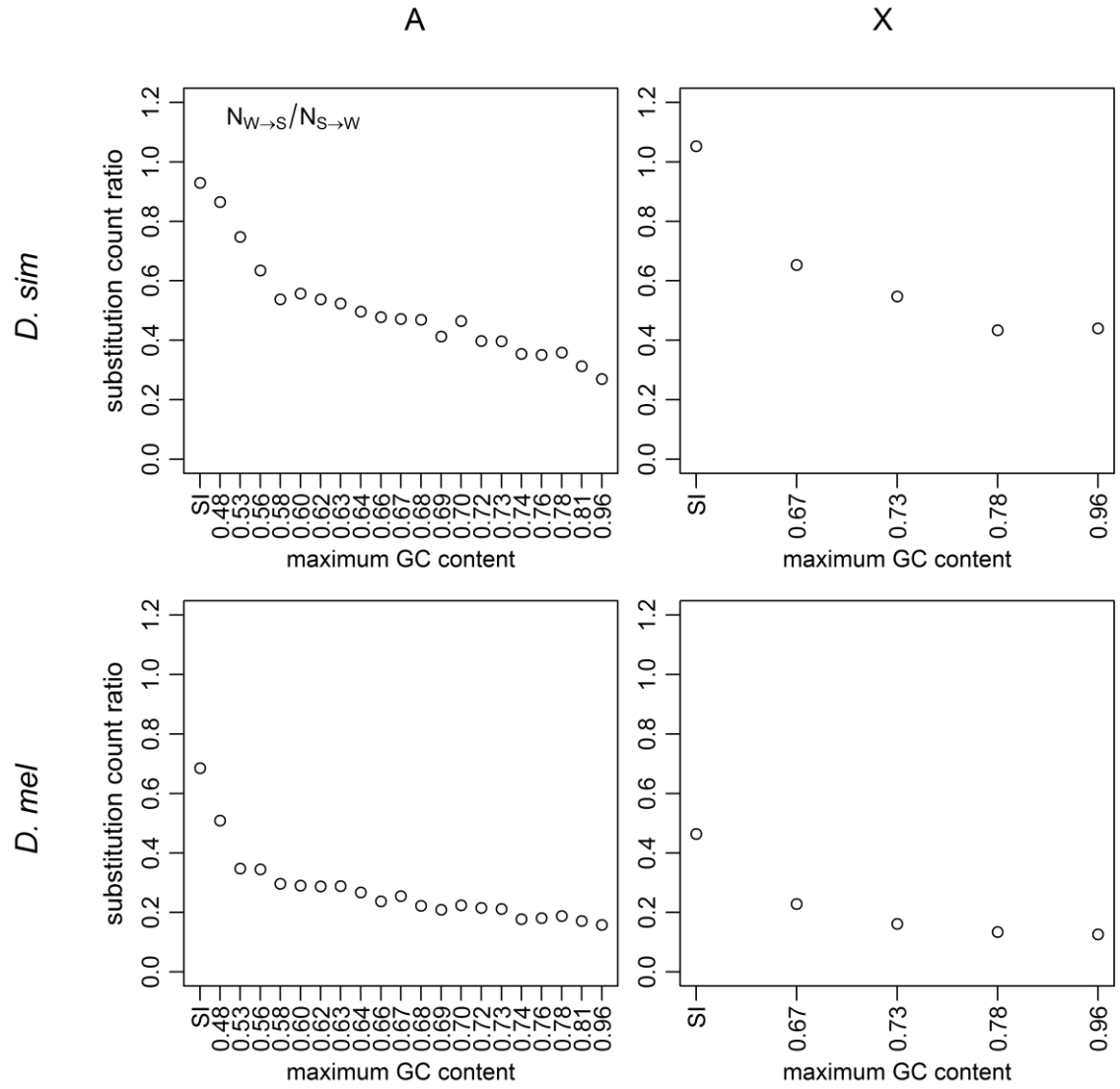


Figure 3. Derived allele frequencies (DAFs) for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant *D. melanogaster* reference sequence. DAF was calculated in the MD (Madagascan) sample of *D. simulans* (top row) and the RG (Rwandan) sample of *D. melanogaster* (bottom row), for autosomes (left-hand column) and X-linked sites (right-hand column). AT → GC mutations – teal circles; GC → AT mutations – orange triangles; AT → AT mutations or GC → GC mutations – lilac squares.

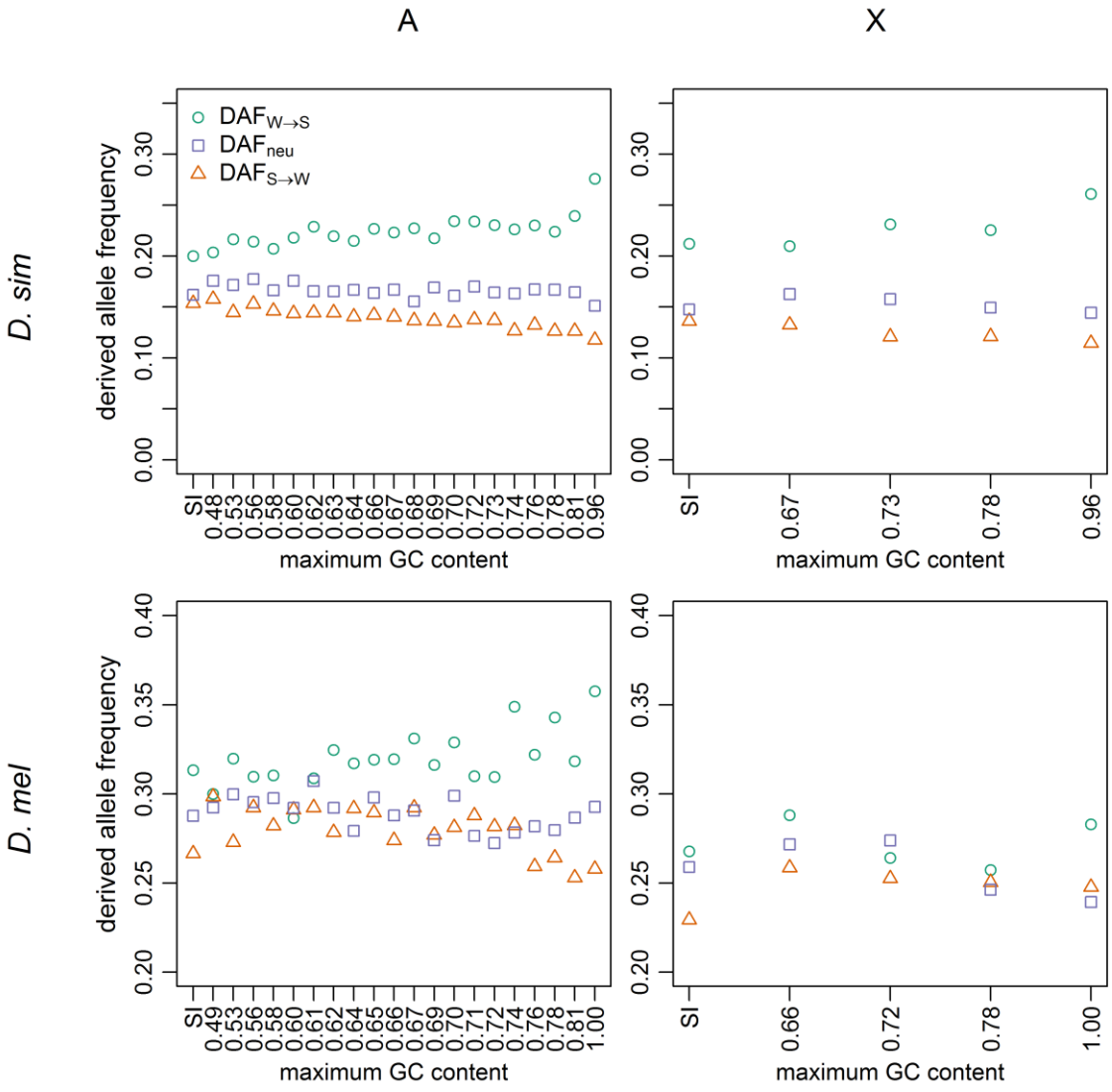
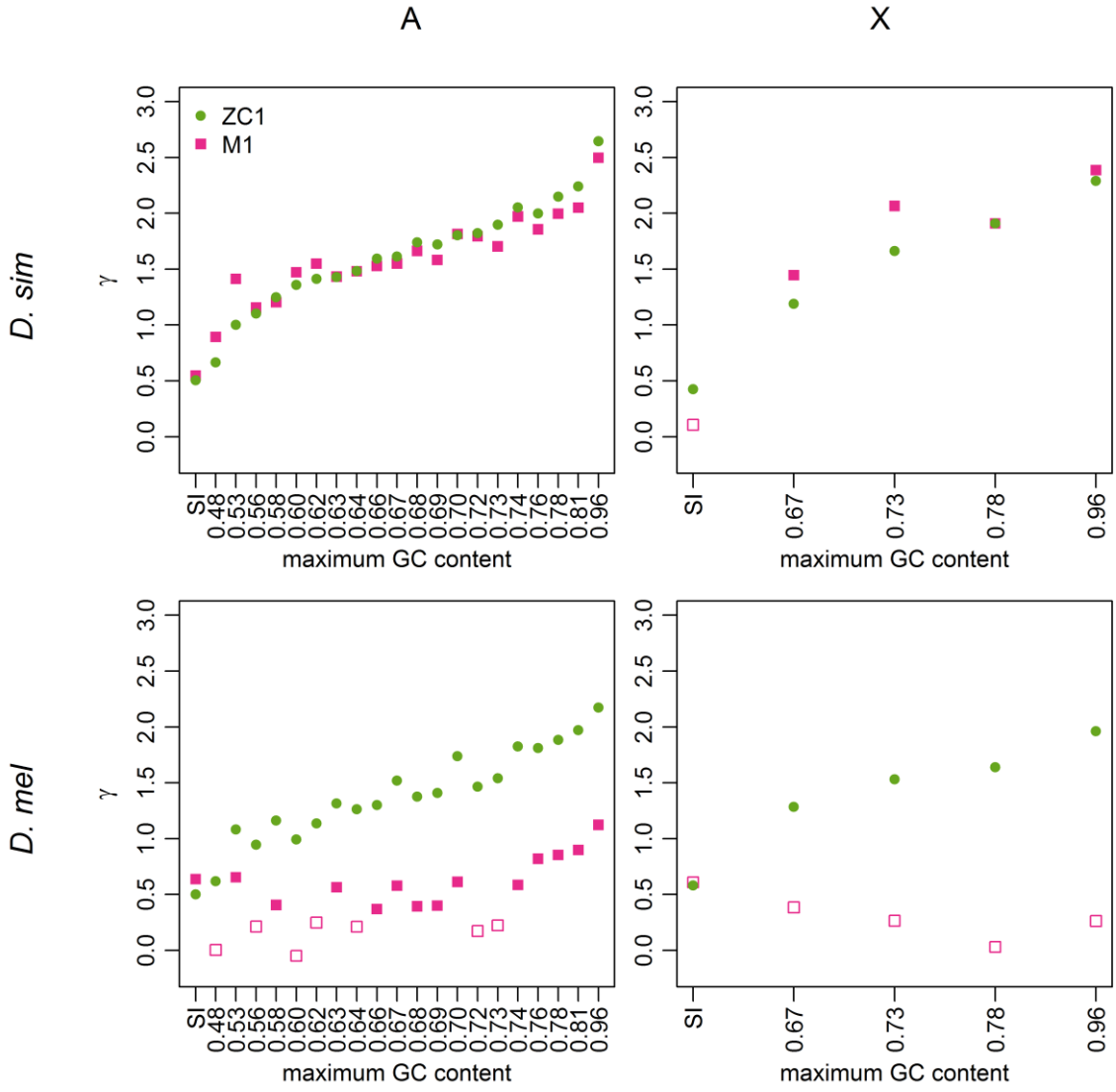


Figure 4. Values of the force favouring GC alleles ($\gamma = 4N_e s$) for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant *D. melanogaster* reference sequence. γ was calculated in the MD (Madagascan) sample of *D. simulans* (top row) and the RG (Rwandan) sample of *D. melanogaster* (bottom row), for autosomes (left-hand column) and X-linked sites (right-hand column), using two methods: the method of Zeng and Charlesworth (2009) with a one-step change in population size (ZC in the main text) – green circles; and the method of Glémin et al (2015) not incorporating polarisation errors (M1 in the main text) – pink squares. Filled points – bins where a model with $\gamma \neq 0$ fitted best; open points – bins where a model with $\gamma = 0$ fitted best.



Supplementary Material

Supplementary text

The difference between the slopes for $r_{S\to W}$ and $r_{W\to S}$

We use the same model and notation as those that lead to Eq. (2) in the main text. As a rough approximation, it suffices to ask whether the ratio of the values of $r_{S\to W}$ observed in the bins with the lowest and highest GC content at 4-fold sites is larger than the ratio for $r_{W\to S}$ between the highest and lowest GC content bins, given the estimated values of γ for the two bins. Consider only the autosomal data. In *D. simulans*, γ was estimated to be 0.28 and 1.27, respectively, for the two extreme GC content bins (see the paragraph below Eq. (3)). The ratio of $r_{S\to W}$ is the ratio of $\gamma/[exp(\gamma)-1]$, whereas the ratio of $r_{W\to S}$ is the ratio of $\gamma/[1-exp(-\gamma)]$. Substituting γ with the two estimates, we can see that, for $r_{S\to W}$, the ratio between the two GC content bins is 1.75, but the ratio is 1.54 for $r_{W\to S}$. The observed values are 2.13 and 1.27, respectively, which differ in the predicted direction.

In *D. melanogaster*, assuming that SI sites are neutrally evolving, then κ can be estimated by dividing $r_{S\to W}$ by $r_{W\to S}$, which turns out to be 2.61. This is somewhat higher than the value of 2, which has been frequently quoted in the literature, but within the range of variation observed in mutation accumulation experiments (Schrider et al. 2013). Using the same approach as above, we first used Eq. (2) in the main text to estimate γ for the two extreme GC bins on the autosomes, which are 0.02 and 1.01, respectively. These values predict that the ratios of $r_{S\to W}$ and $r_{W\to S}$ between the two extreme GC content bins are 1.71 and 1.57. The observed values are 1.93 and 1.40, respectively.

Supplementary Table S1 – Summary statistics for the unfiltered *D. simulans* dataset

Chr[a]	Site	Within population					Between populations
		Pop.[b]	π[c]	θ_W[d]	Δ_π[e]	D[f]	F_{ST}
A	0-fold[g]	MD	0.0017	0.0029	-0.121	-1.330	0.0201
		NS	0.0017	0.0024	-0.094	-0.990	
	4-fold[h]	MD	0.0329	0.0457	-0.081	-1.070	0.0258
		NS	0.0314	0.0380	-0.052	-0.666	
X	0-fold	MD	0.0013	0.0022	-0.124	-1.280	0.0169
		NS	0.0013	0.0018	-0.097	-0.974	
	4-fold	MD	0.0191	0.0296	-0.104	-1.320	0.0242
		NS	0.0182	0.0240	-0.073	-0.890	

All statistics were calculated per gene, and the means are presented here.

[a] Chromosome

[b] Population sample: MD – Madagascar; NS – Kenya

[c] Average number of pairwise differences between lines

[d] Watterson's estimator of θ, the scaled mutation rate

[e] See equation (1)

[f] Tajima's D

[g] 0-fold degenerate sites

[h] 4-fold degenerate sites

Supplementary Table S2 – A comparison between changes along the *D. melanogaster* and *D. simulans* lineages inferred from the dataset presented in this study, and an independent dataset provided by Juraj Bergman and Claud Vogl (pers. comm.). Changes were inferred using parsimony at 4-fold degenerate and SI sites.

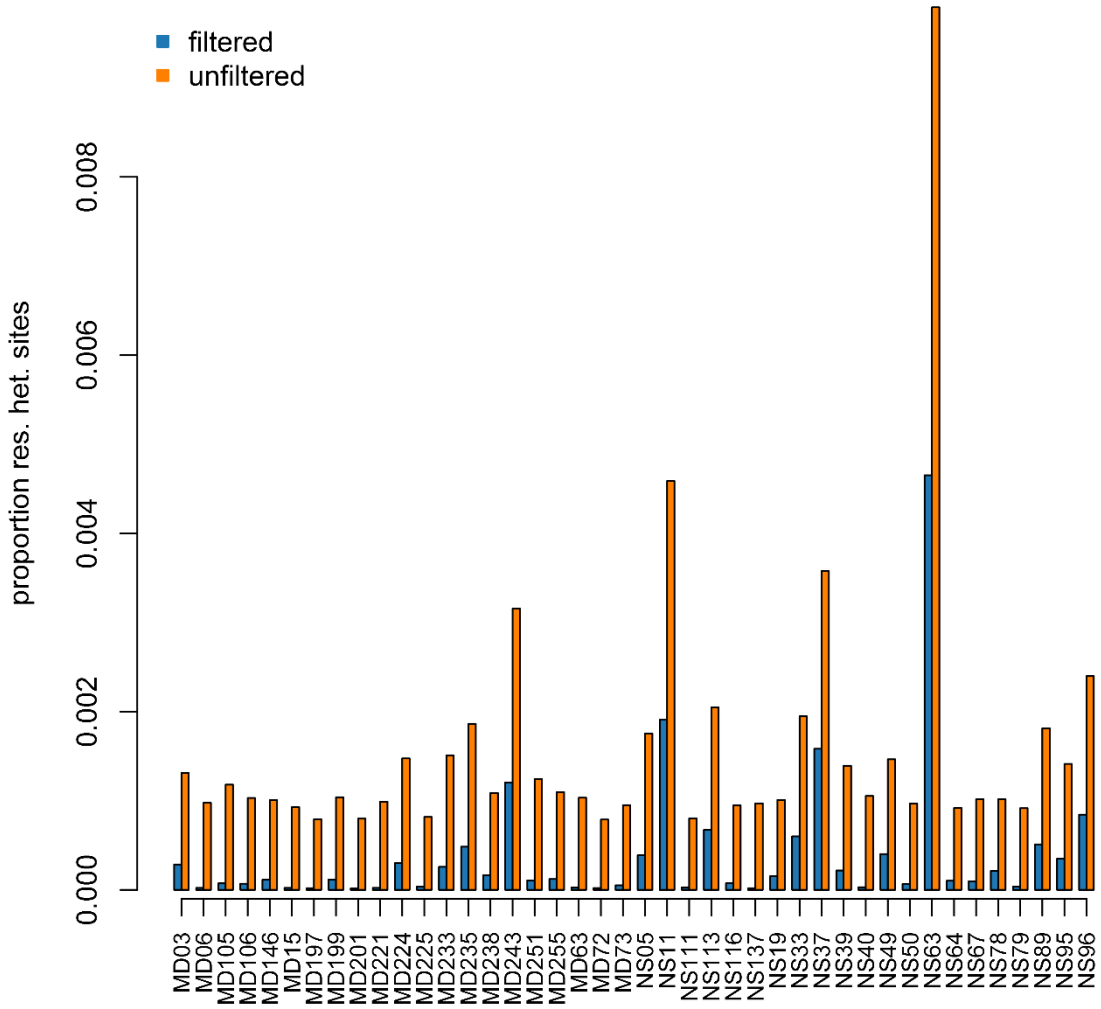
		D. simulans				*D. melanogaster*			
		A		X		A		X	
Site	Dataset	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC	GC → AT
4-fold	This study	13607	25656	1962	3934	10588	40586	1140	7395
	Bergman & Vogl	8178	13917	1581	3789	10142	41858	1219	8587
SI	This study	1859	1598	206	152	1570	1884	131	229
	Bergman & Vogl	1650	1345	196	206	2592	2996	226	352

Supplementary Table S3 – Glémin et al (2015) model AIC rankings. Model M0 - $\gamma = 0$, without polarisation error; Model M1 - $\gamma \neq 0$, without polarisation error; Model M0* - $\gamma = 0$, with polarisation error; Model M1* - $\gamma \neq 0$, with polarisation error.

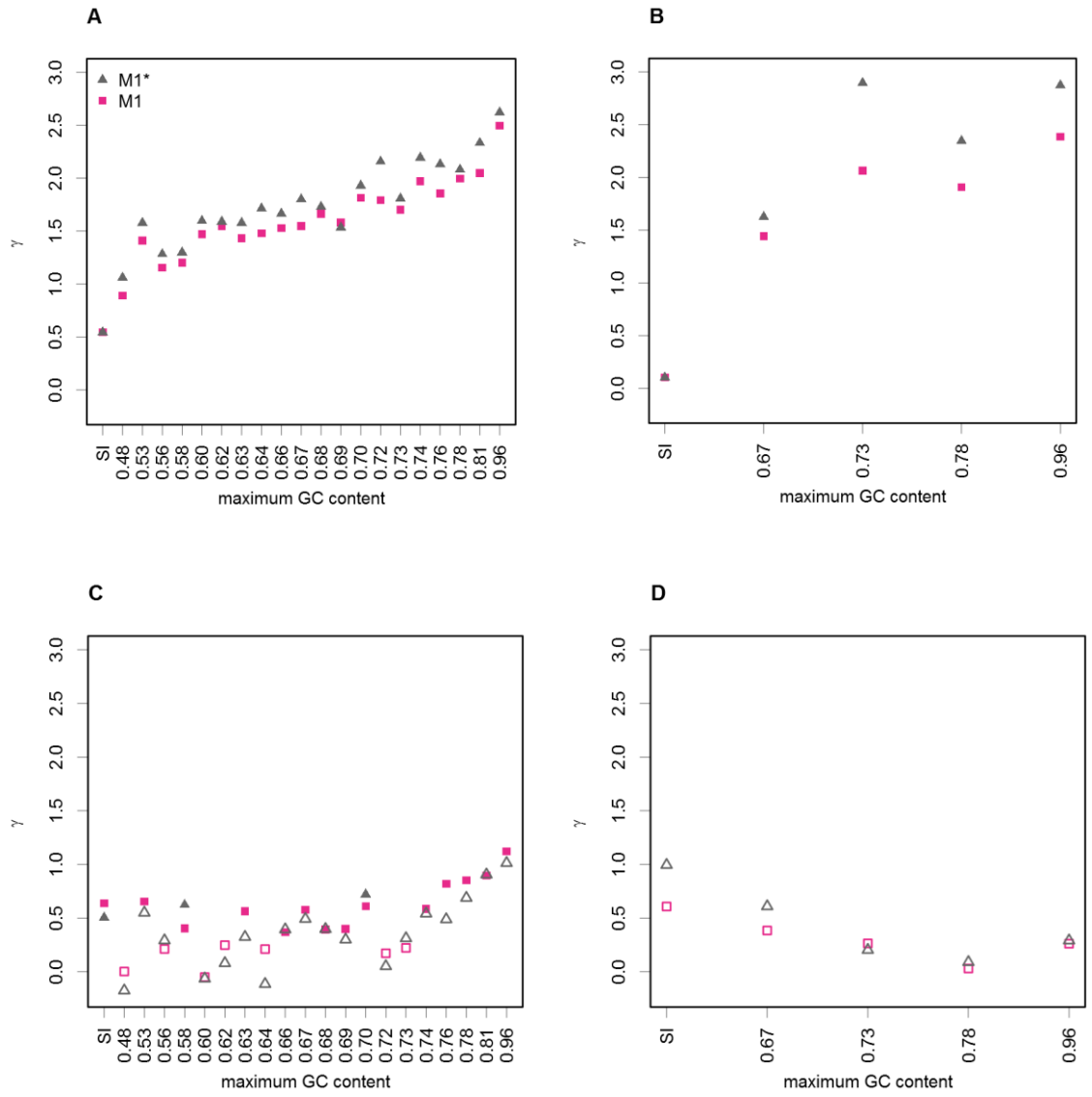
				AIC model ranking (1 is the best)			
Species	Chr.	Site	Bin	M0	M0*	M1	M1*
simulans	A	SI		3	4	1	2
		4-fold	1	4	3	1	2
			2	4	3	1	2
			3	4	3	1	2
			4	4	3	2	1
			5	4	3	1	2
			6	4	3	2	1
			7	4	3	1	2
			8	4	3	2	1
			9	4	3	1	2
			10	4	3	2	1
			11	4	3	2	1
			12	4	3	1	2
			13	4	3	2	1
			14	4	3	2	1
			15	4	3	2	1
			16	4	3	1	2
			17	4	3	2	1
			18	4	3	1	2
			19	4	3	2	1
			20	4	3	2	1
	X	SI		1	3	2	4
		4-fold	1	4	3	1	2
			2	4	3	2	1
			3	4	3	2	1
			4	4	3	2	1
melanogaster	A	SI		4	3	1	2
		4-fold	1	1	3	2	4
			2	4	3	1	2
			3	1	3	2	4
			4	2	4	1	3
			5	1	3	2	4

		6	1	3	2	4
		7	4	2	1	3
		8	1	3	2	4
		9	2	3	1	4
		10	4	2	1	3
		11	4	2	1	3
		12	2	3	1	4
		13	3	4	1	2
		14	1	3	2	4
		15	1	3	2	4
		16	2	3	1	4
		17	4	2	1	3
		18	4	2	1	3
		19	4	3	1	2
		20	4	3	1	2
	X	SI	2	4	1	3
		4-fold 1	2	4	1	3
		2	1	3	2	4
		3	1	3	2	4
		4	1	3	2	4
	A	4-fold 1&2	3	2	1	4
		3&4	3	4	1	2
		5&6	1	3	2	4
		7&8	4	2	1	3
		9&10	3	4	1	2
		11&12	4	2	1	3
		13&14	2	4	1	3
		15&16	2	4	1	3
		17&18	4	3	1	2
		19&20	4	3	1	2

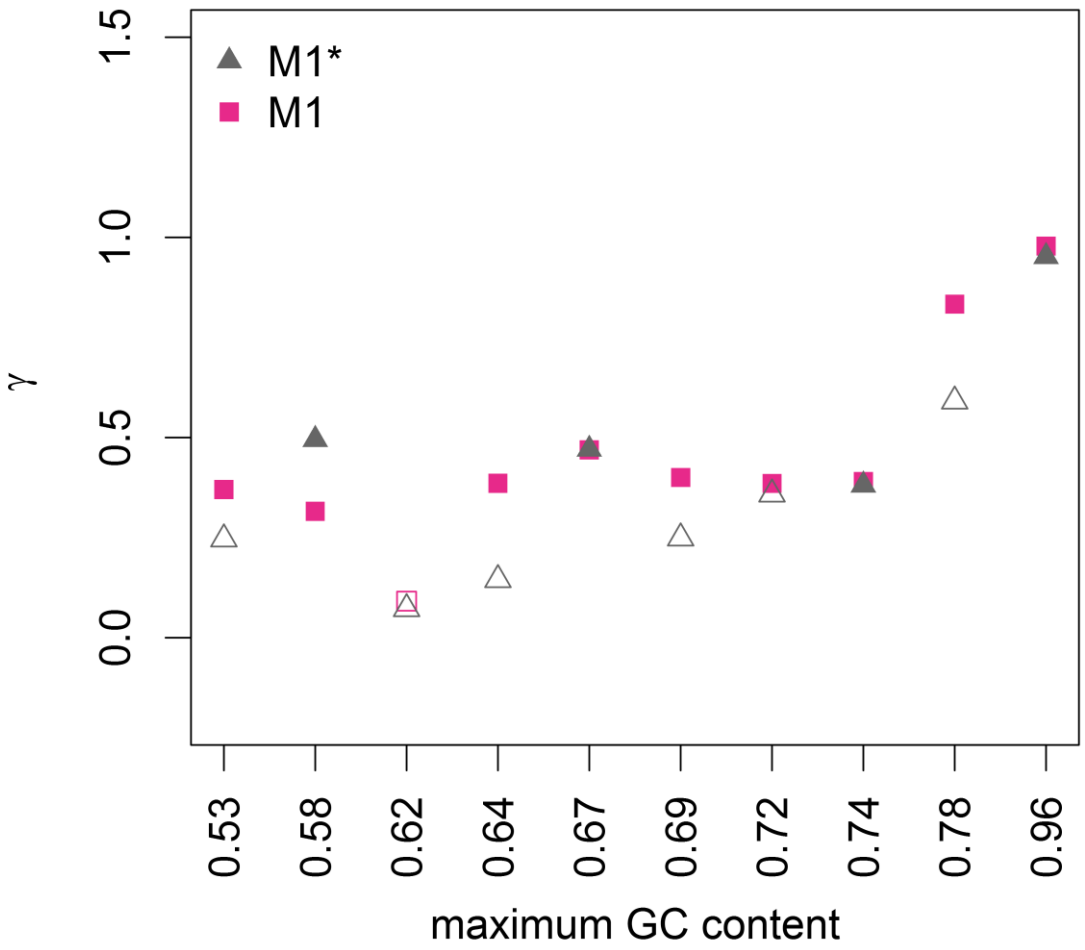
Supplementary Figure S1. **Residual heterozygosity per *D. simulans* isofemale line, calculated as the proportion of sites that were called as heterozygotes in the entire genome.** Orange columns – the unfiltered dataset; blue columns – the 95% VQSR filtered dataset (see Methods).



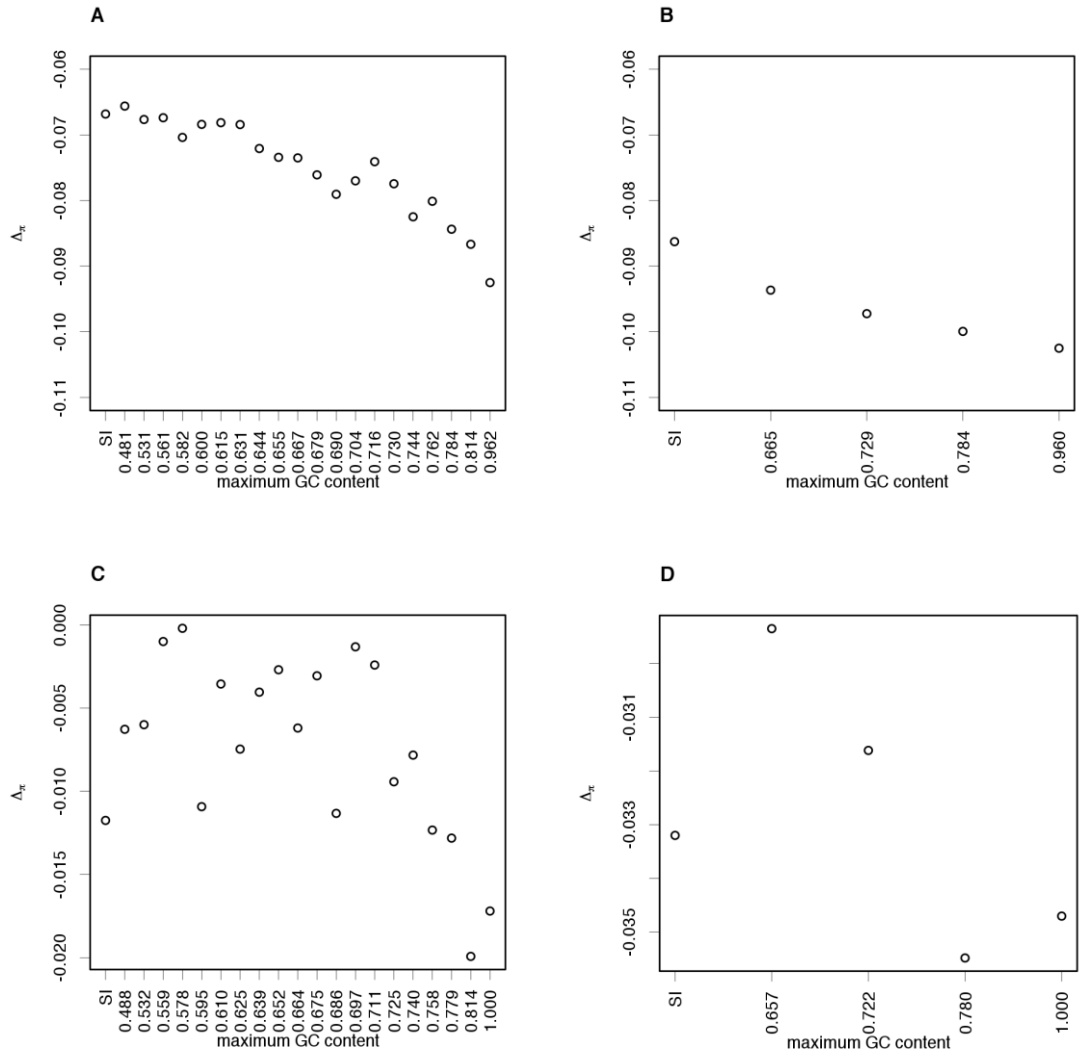
Supplementary Figure S2. Comparison between estimates of values of the force favouring GC alleles ($\gamma = 4N_e s$) from the methods of Glémin et al (2015). Panel A – autosomal sites from the Madagascan (MD) *D. simulans* sample; panel B – X-linked sites from the MD *D. simulans* sample. Panel C – autosomal sites from the Rwandan (RG) *D. melanogaster* sample; panel D – X-linked sites from the RG *D. melanogaster* sample. Pink squares – model M1; grey triangles – model M1*. Filled points – bins where a model with $\gamma \neq 0$ fitted best; open points – bins where a model with $\gamma = 0$ fitted best.



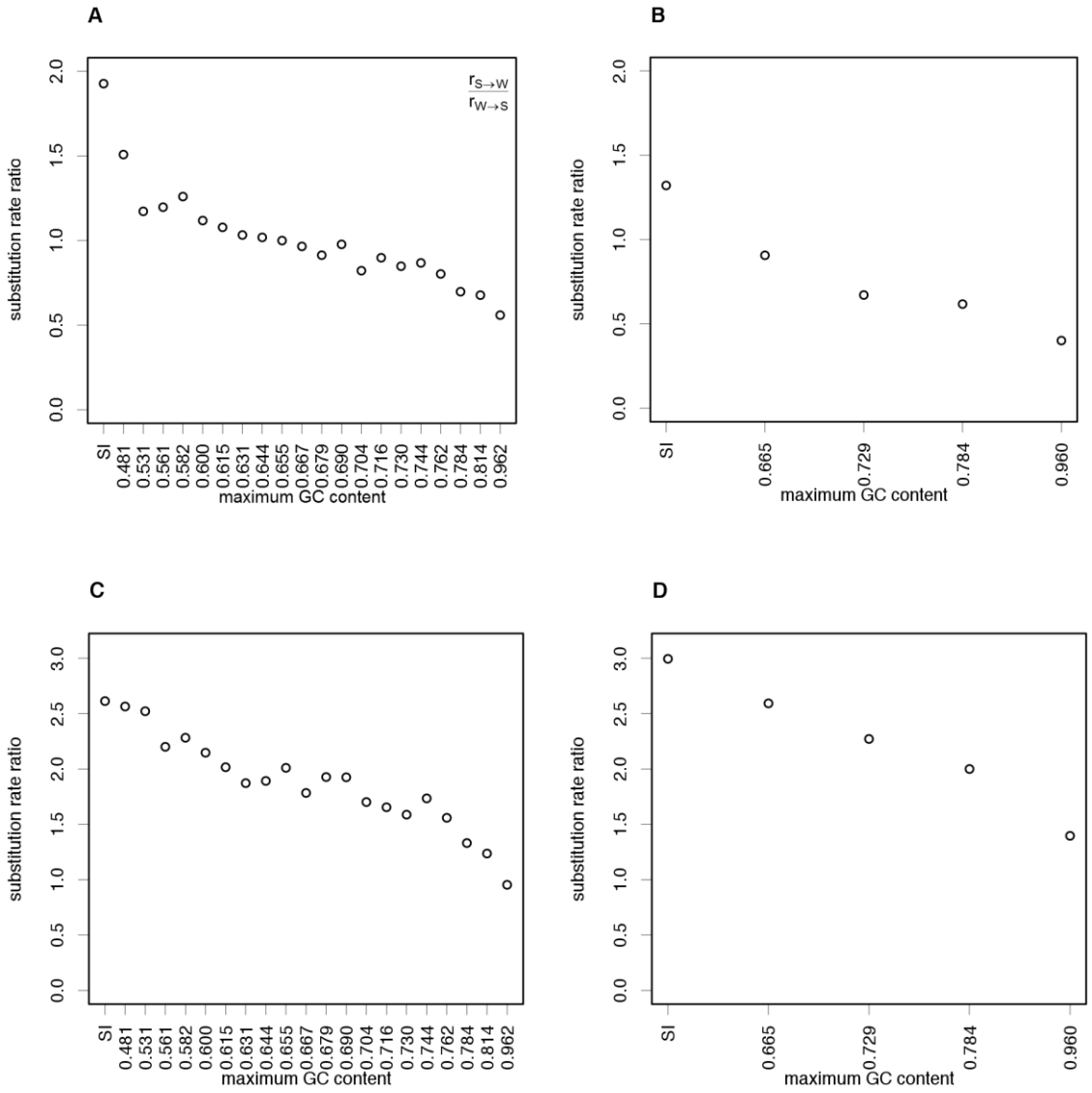
Supplementary Figure S3. **Estimates of values of the force favouring GC alleles ($\gamma = 4N_e s$) for Rwandan (RG) *D. melanogaster* autosomal 4-fold degenerate sites, reduced to 10 GC content bins.** γ was calculated using the method of Glémin et al (2015) incorporating polarisation errors (M1* in the main text) – grey triangles; and not incorporating polarisation errors (M1 in the main text) – pink squares. Filled points – bins where a model with $\gamma \neq 0$ fitted best; open points – bins where a model with $\gamma = 0$ fitted best.



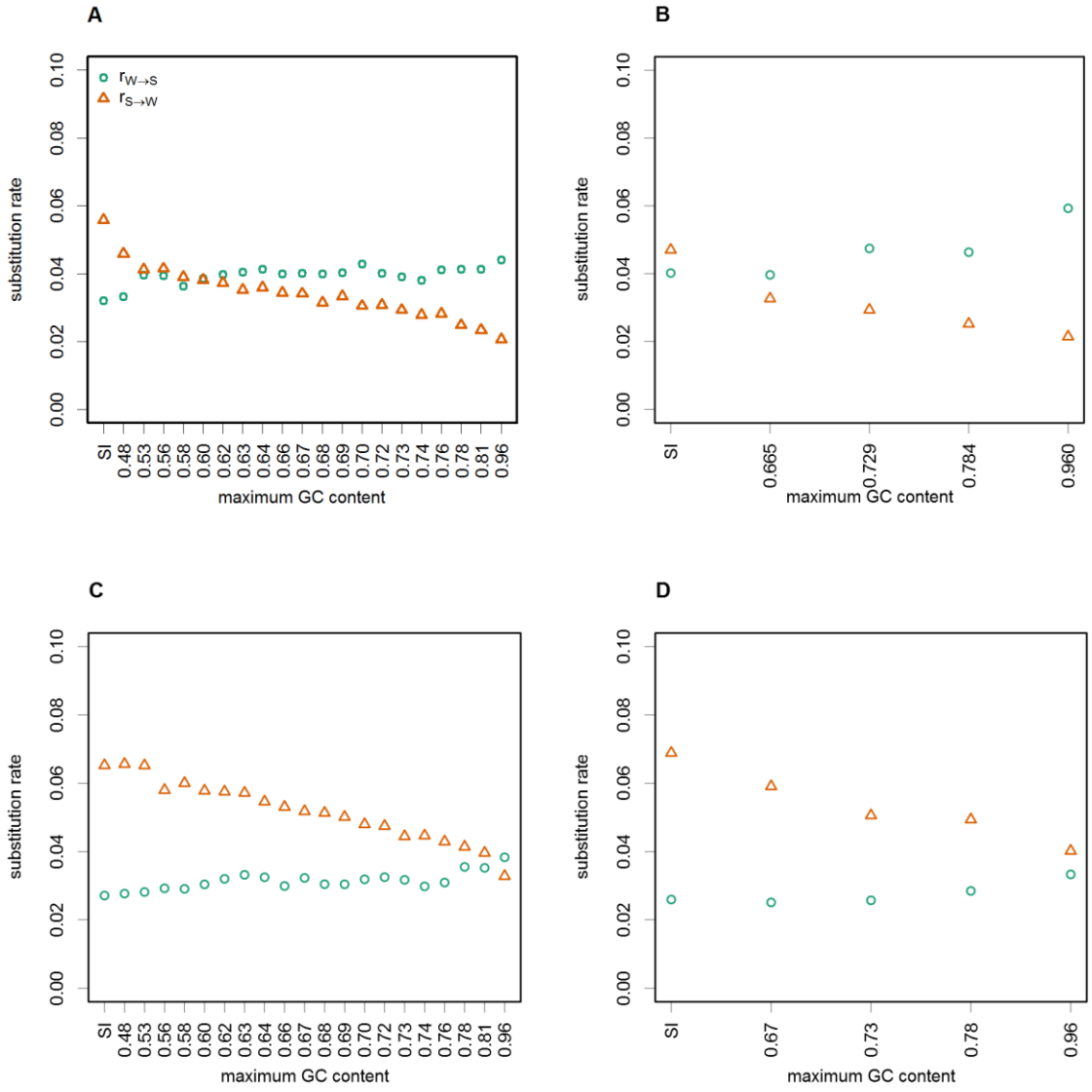
Supplementary Figure S4. The Δ_π statistic of Langley et al (2014). Panel A – autosomal sites from the Madagascan (MD) *D. simulans* sample; panel B – X-linked sites from the MD *D. simulans* sample; Panel C – autosomal sites from the Rwandan (RG) *D. melanogaster* sample; panel D – X-linked sites from the RG *D. melanogaster* sample. In *D. simulans*, Δ_π is negatively correlated with GC content at autosomal sites (Kendall's τ = -0.88, p < 0.001; panel A). In *D. melanogaster*, Δ_π is also negatively correlated with GC content at autosomal sites (Kendall's τ = -0.41, p = 0.012; panel C).



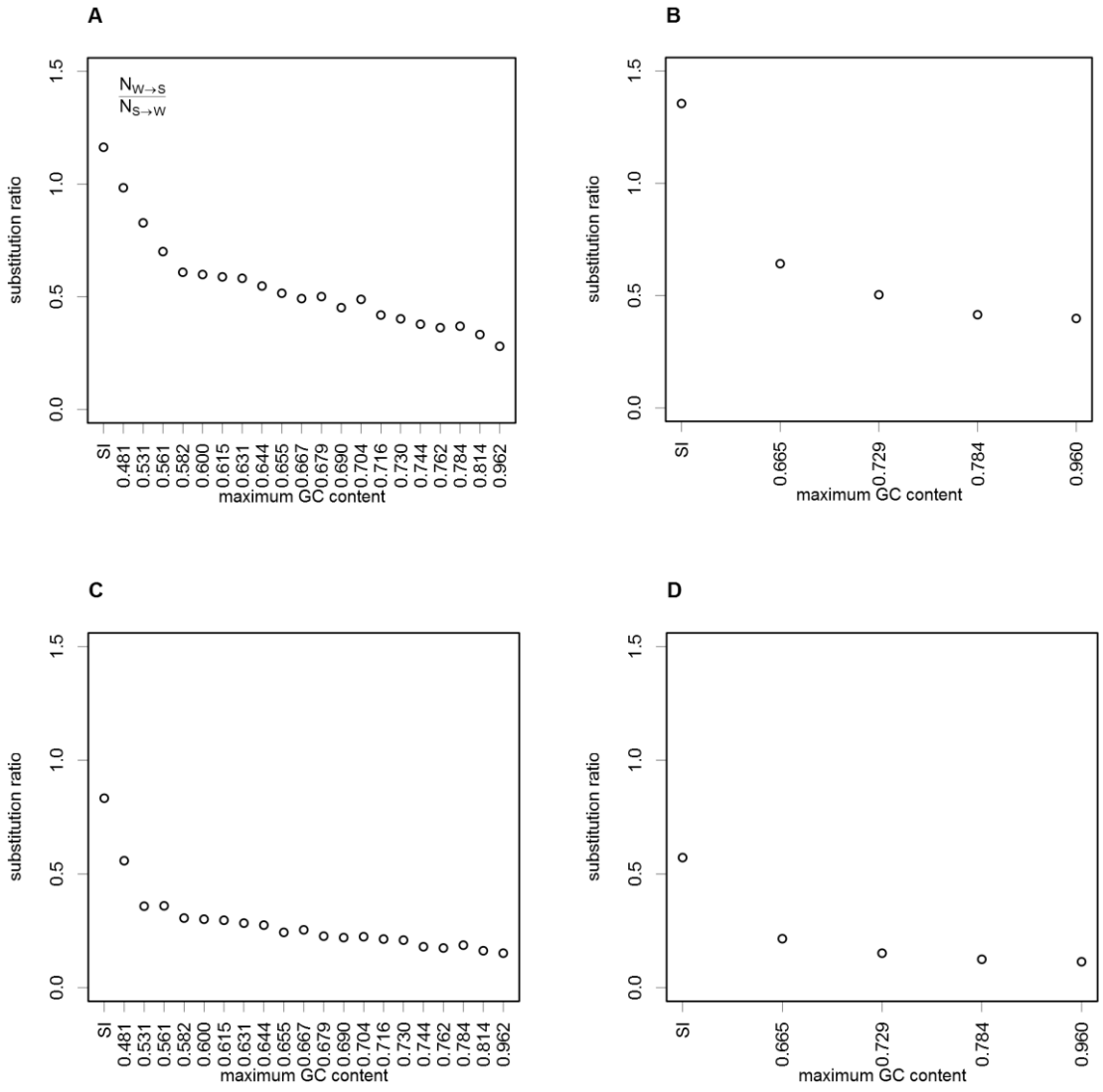
Supplementary Figure S5. The ratio of substitution rates for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant *D. melanogaster* reference sequence, and using AWP to infer ancestral states. Panel A – autosomal substitution rates ratios along the *D. simulans* lineage; panel B – X-linked substitution rate ratios along the *D. simulans* lineage; panel C – autosomal substitution rate ratios along the *D. melanogaster* lineage; panel D – X-linked substitution rate ratios along the *D. melanogaster* lineage.



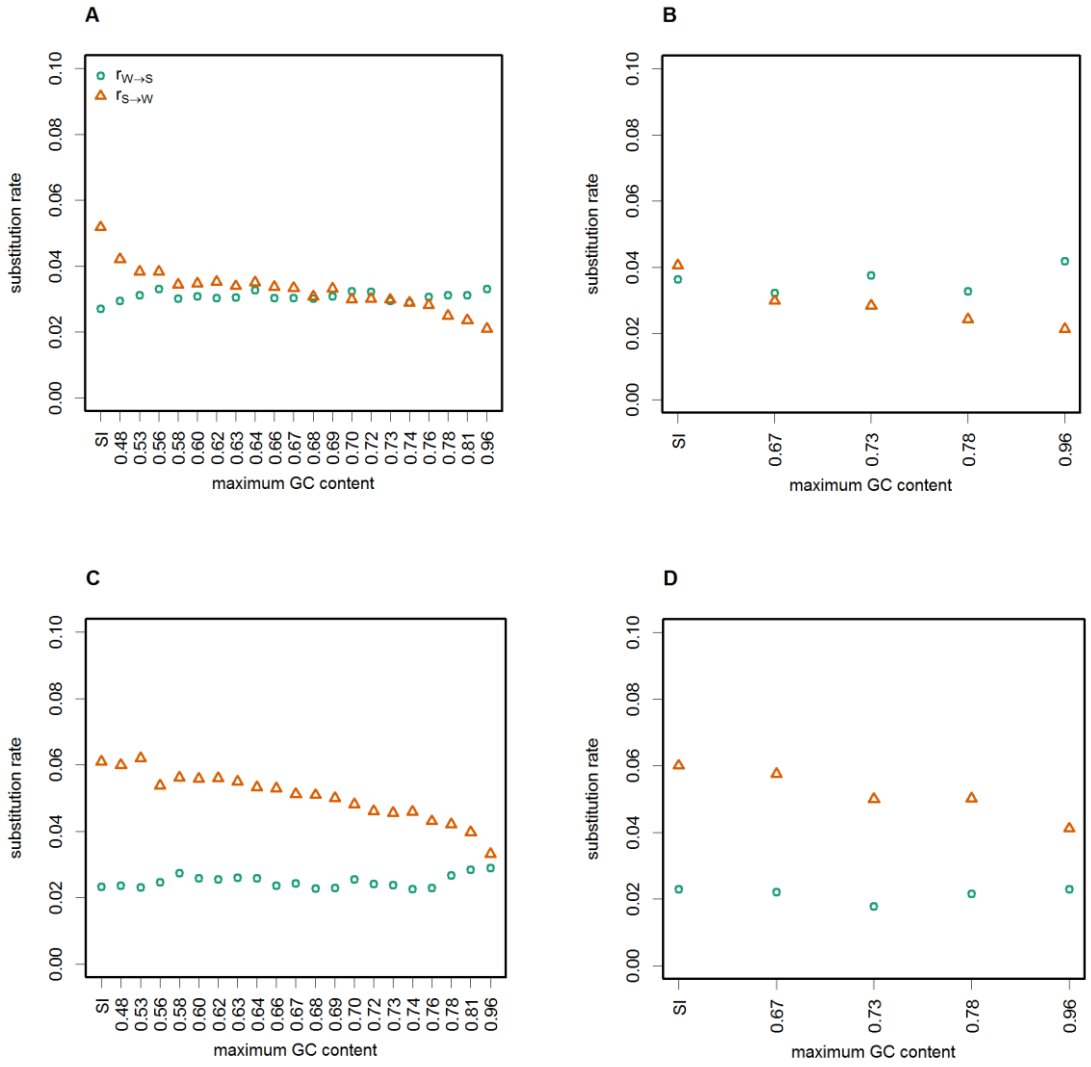
Supplementary Figure S6 – Substitution rates for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant D. melanogaster reference sequence, for AT → GC substitutions (teal circles) and GC → AT substitutions (orange triangles) using parsimony to infer ancestral states. Panel A – autosomal substitution rates along the *D. simulans* lineage; panel B – X-linked substitution rates along the *D. simulans* lineage; panel C – autosomal substitution rates along the *D. melanogaster* lineage; panel D – X-linked substitution rates along the *D. melanogaster* lineage.



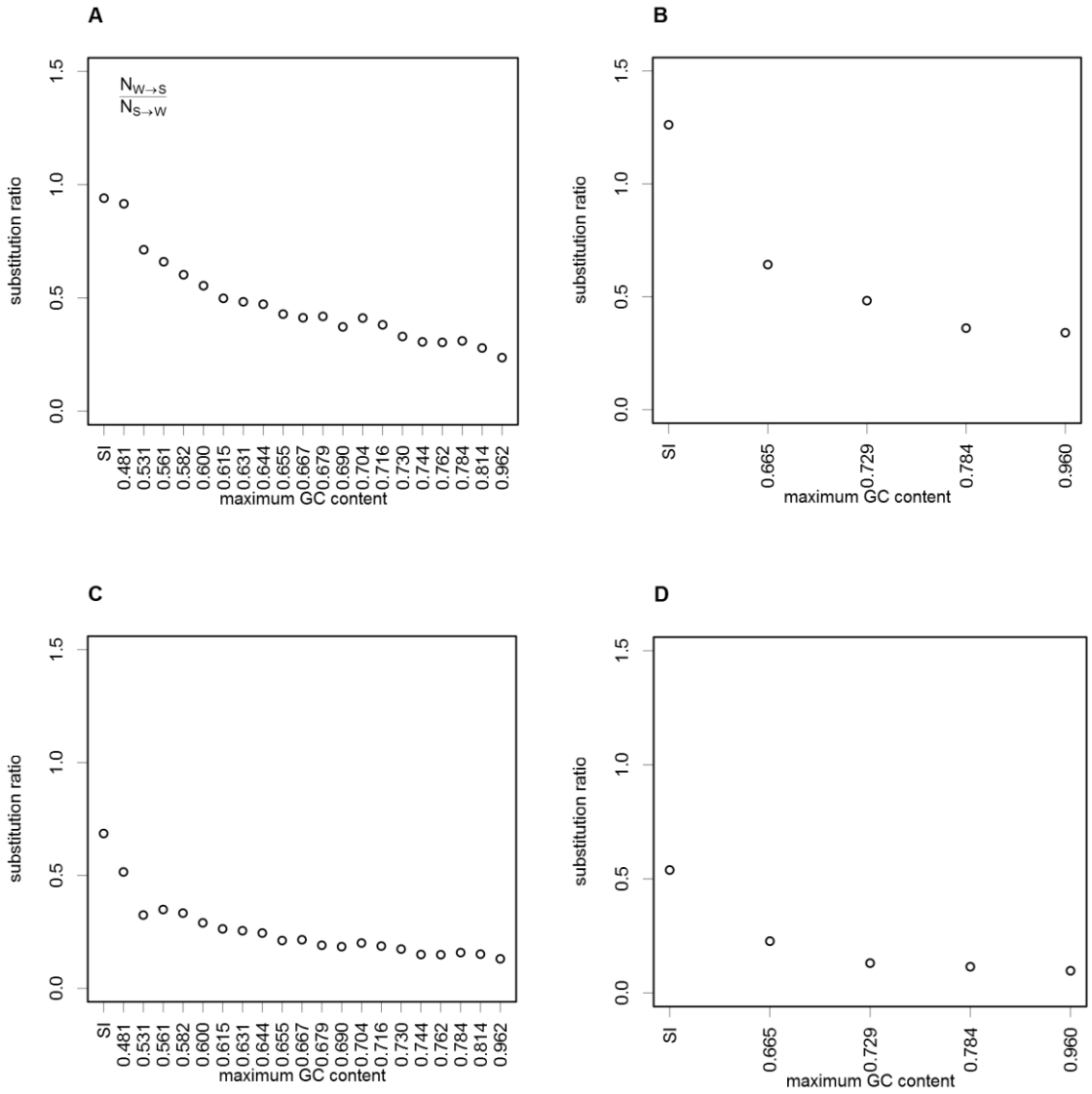
Supplementary Figure S7 – The ratio of substitution counts for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant *D. melanogaster* reference sequence and using parsimony to infer ancestral states. Panel A – autosomal ratio of substitution counts along the *D. simulans* lineage; panel B – X-linked ratio of substitution counts along the *D. simulans* lineage; panel C – autosomal ratio of substitution counts along the *D. melanogaster* lineage; panel D – X-linked ratio of substitution counts along the *D. melanogaster* lineage.



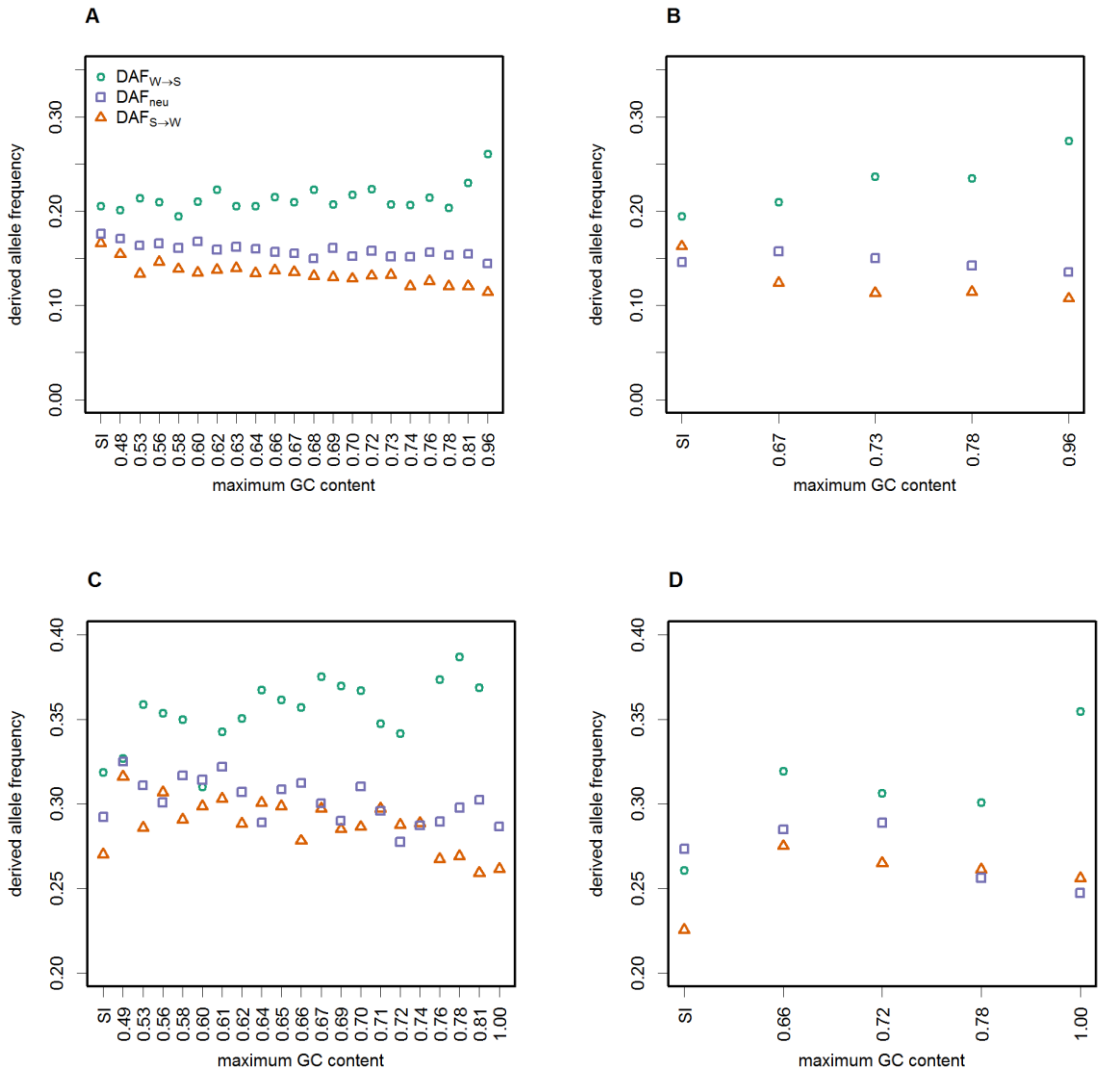
Supplementary Figure S8 – Substitution rates for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant *D. melanogaster* reference sequence, for AT → GC substitutions (teal circles) and GC → AT substitutions (orange triangles) using SBP to infer ancestral states. Panel A – autosomal substitution rates along the *D. simulans* lineage; panel B – X-linked substitution rates along the *D. simulans* lineage; panel C – autosomal substitution rates along the *D. melanogaster* lineage; panel D – X-linked substitution rates along the *D. melanogaster* lineage.



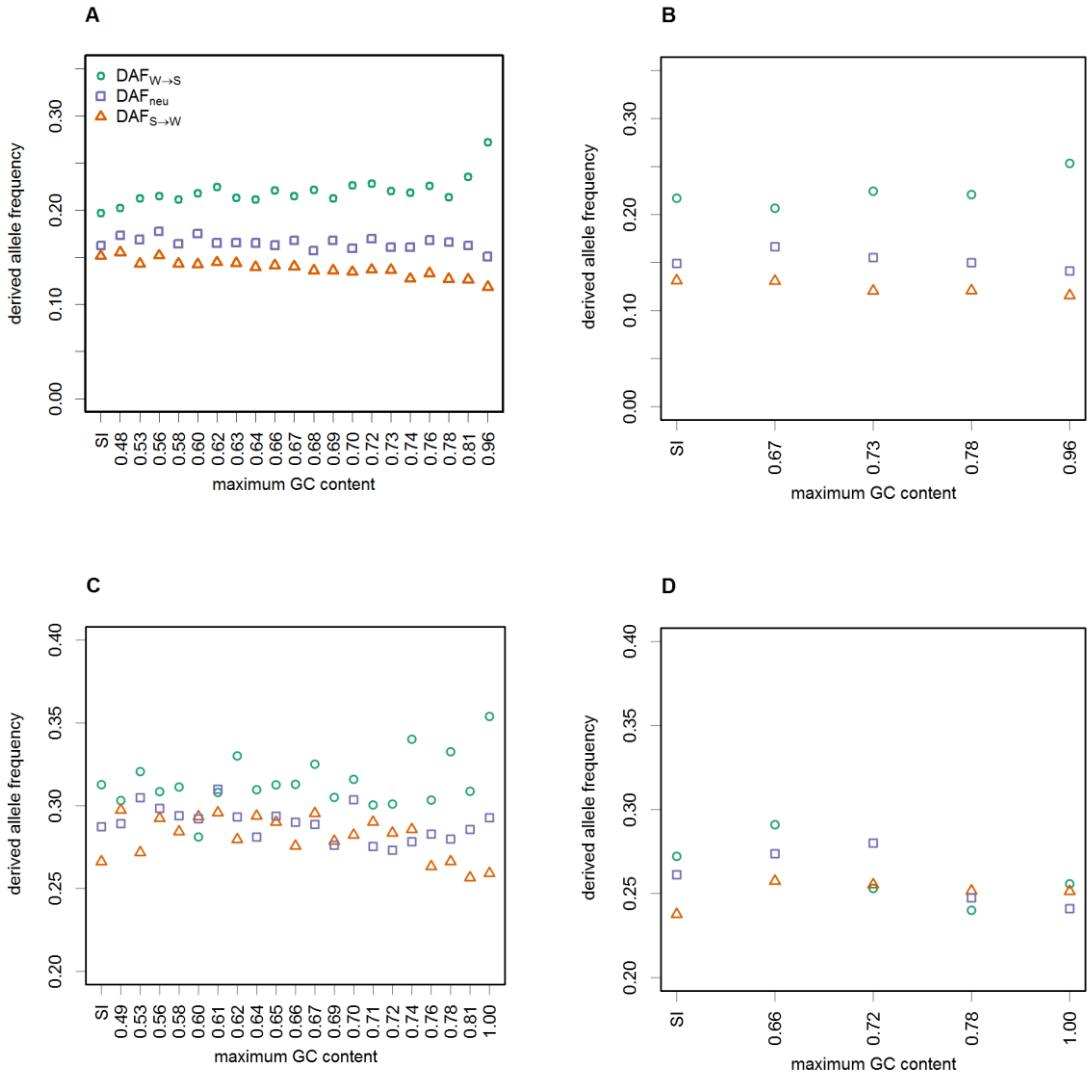
Supplementary Figure S9 – The ratio of substitution counts for positions 8-30bp of introns <66bp long (SI sites), and 4-fold degenerate sites binned by the GC content of the extant *D. melanogaster* reference sequence and using SBR to infer ancestral states. Panel A – autosomal ratio of substitution counts along the *D. simulans* lineage; panel B – X-linked ratio of substitution counts along the *D. simulans* lineage; panel C – autosomal ratio of substitution counts along the *D. melanogaster* lineage; panel D – X-linked ratio of substitution counts along the *D. melanogaster* lineage.



Supplementary Figure S10 – Derived allele frequencies for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant *D. melanogaster* reference sequence, and using parsimony to infer ancestral states. AT → GC mutations – teal circles; GC → AT mutations – orange triangles; AT → AT mutations or GC → GC mutations – lilac squares. Panel A – autosomal DAFs in the Madagascan (MD) *D. simulans* sample; panel B – X-linked DAFs in the MD *D. simulans* sample; panel C – autosomal DAFs in the Rwandan (RG) *D. melanogaster* sample; panel D – X-linked DAFs in the RG *D. melanogaster* sample.



Supplementary Figure S11 – Derived allele frequencies for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points) binned by the GC content of the extant *D. melanogaster* reference sequence, and using SBR to infer ancestral states. AT → GC mutations – teal circles; GC → AT mutations – orange triangles; AT → AT mutations or GC → GC mutations – lilac squares. Panel A – autosomal DAFs in the Madagascan (MD) *D. simulans* sample; panel B – X-linked DAFs in the MD *D. simulans* sample; panel C – autosomal DAFs in the Rwandan (RG) *D. melanogaster* sample; panel D – X-linked DAFs in the RG *D. melanogaster* sample.



Chapter 5. General Conclusions

In the previous three chapters, I have tried to understand the forces that affect genomic evolution in natural populations. These forces include natural selection (of different types), mutation, GC-biased gene conversion, genetic drift and demography. As well as being various, they are not mutually exclusive, and the action of one may confound the detection of the others (see Chapter 1). To tease apart which forces are acting and to what degree, I have used a comparative method between different classes of sites, and between different genomic contexts, as well as by using both divergence and polymorphism data. Where appropriate, I have explicitly incorporated demography into these analyses (Chapter 3), or have used methods that account for non-equilibrium when inferring other population genetic parameters (Chapter 4). In Chapter 2, I investigated the effect that selective forces have on measures of demography (i.e. population subdivision), rather than the emphasis being the other way around. Below, I briefly summarise some of the ways in which these different forces bear on the results from the three chapters.

Non-equilibrium

If the forces of mutation, selection and drift do not balance each other then the composition of genomes may change over time in non-random ways. Such departures from equilibrium may be caused by demographic events. For instance, a recent population expansion (without a bottleneck) may mimic the signal of background selection in a population of moderate size because both are expected to cause an excess of low-frequency variants compared to the standard neutral expectation (Charlesworth et al. 1993). A bottleneck may mimic the effect of a partial hitchhiking event (i.e. a selective sweep with some recombination), because both are expected to result in a reduction of low-frequency variants compared to the neutral expectation (Charlesworth et al. 1993).

In Chapter 3, I inferred a population expansion in Gouldian Finches using putatively neutrally evolving sites in randomly chosen Z-linked introns. I incorporated this information into the coalescent simulations that I used as a test of neutral evolution at

the *Red* locus, in order to control for the effects of demography when trying to infer whether selection has acted at the *Red* locus. I also carried out the same procedures without incorporating the inferred expansion, which, in the most part, returned similar results. However, the extent of the deviation from the neutral expectation is generally smaller under the constant size model (c.f. Figure 4 and Supplementary Figure S9 (panels A-D) in Chapter 3), which serves to highlight the importance of correctly accounting for demography when trying to infer other evolutionary parameters.

In Chapter 4, part of the motivation behind the analysis of codon usage bias (CUB) in *Drosophila* was that previous work in *D. melanogaster* was complicated by the fact that its GC content is not at equilibrium (Akashi 1995; Akashi 1996; Poh et al. 2012). This has implications for inferring ancestral states (Eyre-Walker 1998; Matsumoto et al. 2015), a process that is important for interpreting polarised divergence and polymorphism data. Various explanations for the non-equilibrium in *D. melanogaster* have been proposed, one of which is a historic reduction in population size (Akashi 1996), which would have the effect of weakening the scaled strength of selection for preferred codons, resulting in a shift towards AT at synonymous sites. I wanted to revisit this topic using data from *D. simulans*, which we hoped might be at base content equilibrium, although I found that this was not the case (Tables 2 and 3 in Chapter 4). In order to deal with the potentially misleading effects of demographic changes (or other factors; see below) in both *D. simulans* and *D. melanogaster*, we therefore decided to use methods that account for the effects of non-equilibrium when inferring evolutionary parameters (when polarising segregating sites and estimating the scaled strength of selection for GC alleles), in order to mitigate the potentially confounding effects of demographic changes. This methodology allowed us to resolve some previously conflicting reports about the evolution of CUB in *Drosophila* (see below).

Confounding deterministic factors

Different confounding factors are considered at times in this thesis. For instance, GC-biased gene conversion (gBGC) and selection for preferred codons are both expected to increase the GC content of *Drosophila* genomes. Conversely, an increase in the mutational bias towards AT, or a decrease in the scaled strength of selection for preferred codons (perhaps due to a reduction in N_e) are both expected to decrease it.

Changes in the strength of these forces are expected to lead to a period of time in which genome composition is not at equilibrium.

GC-biased gene conversion affects substitution rates as well as the site frequency spectrum in a way that can mimic selection (Duret and Galtier 2009). Evidence for gBGC in nature is available mainly from microbes and mammals, but it has recently been shown to play a role in the evolution of bird genomes (Bolívar et al. 2015). The evidence for gBGC from intronic sequences in Gouldian Finches in Chapter 3 of this thesis provides further evidence in support of this. In Chapter 4, I wanted to discount the effects of gBGC when quantifying the extent of selection for preferred codons in *Drosophila*. To do so, I used sequences from short introns (SIs), which are expected to experience gBGC, but not selection for preferred codons, to control for its effects. Previous work has suggested that there is little evidence for gBGC in *Drosophila melanogaster* (Clemente & Vogl 2012b; Comeron et al. 2012; Campos et al. 2013; Robinson et al. 2014), although there is some evidence for a GC-favouring force in X-linked non-coding sequences in *D. simulans* (Haddrill and Charlesworth 2008). Contrary then, to our expectation not to find any evidence of gBGC in Chapter 4, there was some evidence of a force that favours GC in short introns in both *D. simulans* and *D. melanogaster*. This was apparent from derived allele frequencies (DAFs) at short introns – where AT-ancestral, GC-derived variants segregated at higher frequencies than GC-ancestral, AT-derived variants (Figure 3 in Chapter 4) – as well as from our estimates of $\gamma = 4N_e s$, where s is the selection intensity favouring GC alleles, which were significantly greater than 0 in autosomal SIs in both species (Figure 4 in Chapter 4). However, in the absence of more detailed recombination information in *D. simulans*, whether these signals are attributable to gBGC, or some other process, remains to be seen. Regardless, gBGC alone is unlikely to be able to explain the patterns we observed at 4-fold degenerate sites, given that our estimates of γ at those sites are consistently higher than in SIs, as well as previous evidence which is consistent with the action of selection for preferred codons, but not gBGC in *D. melanogaster* (Campos et al. 2013).

Single nucleotide mutation events are AT-biased in most, if not all, taxa studied to date (Schrider et al. 2013 and references therein), including the Gouldian Finch (Chapter 3). This phenomenon bears on our analysis of CUB in Chapter 4 in particular, because an alternative explanation for the non-equilibrium base composition in *D. melanogaster* to

a reduction in effective population size is that there has been a shift in mutational bias in *D. melanogaster*, further towards AT. To distinguish these two possibilities, I compared the bias towards GC across different regions of the genome – the ratio of $AT \rightarrow GC/GC \rightarrow AT$ substitution counts correlated negatively with GC content (Figure 2 in Chapter 4), which could most parsimoniously be explained by a reduction in the scaled strength of selection for preferred codons. The alternative is that the mutational bias has shifted to differing degrees in different GC content regions, which was judged unlikely, as well as not being able to explain other features of *Drosophila* genome evolution (Campos et al. 2013; Chapter 4).

Selection

In Chapter 2, I investigated the extent to which natural selection affects genetic differentiation between French and Rwandan populations of *D. melanogaster*. The original rationale for this study was to try to understand how linkage to selected sites might influence genome-wide patterns of F_{ST}. Because F_{ST} is the ratio of subpopulation diversity to total diversity, forces that affect either of these components also affect F_{ST}, which makes it hard to interpret. We found that 0-fold F_{ST} *did* correlate strongly (negatively) with the level of constraint (Figure 2 in Chapter 2). Our first explanation for this pattern was that faster evolving genes might be more likely to be involved in local adaptation (to one or both of the French and Rwandan environments). However, a closer examination of the data revealed that this relationship could be better explained by F_{ST}'s dependence on minor allele frequency. 0-fold sites in genes under higher constraint experience stronger purifying selection, which in turns keeps the frequency of deleterious mutations lower, and this results in a reduced upper bound of the value which F_{ST} can take. There has been some difficulty in linking population genetics-based signatures of selection, including those based on F_{ST}, to the actual targets of selection (Akey 2009). The results from this chapter serve to highlight now understanding its statistical properties might help in interpreting biological processes.

In Chapter 3, patterns of genetic diversity at the *Red* locus were found to be compatible with natural selection maintaining a long-term, balanced polymorphism. Specifically, diversity within the two alleles at the *Red* locus (*b* and *R*) was (generally) similar to the wider genetic background on the Z chromosome. When combining information from

both alleles however, patterns of polymorphism were significantly different from the surrounding genetic regions: F_{ST} between *b* and *R* was nearly maximal, and nucleotide diversity including sequences from both *b* and *R* alleles was much elevated over the neutral expectation, as was Tajima's *D* (Table 1 and Figure 3 in Chapter 3). Although on the face of it, these patterns are consistent with balancing selection, the sampling scheme employed at the *Red* locus was such that these patterns were expected, to some extent, even if the *Red* locus has been evolving neutrally. As a consequence of this, it was necessary to take the sampling scheme into account to determine whether the observed data required the action of a force (namely, selection) over and above the effect of non-random sampling alone. I found that they did (Figures 4 and 5 in Chapter 3). Although the reasons for the maintenance of two distinct colour pattern alleles at the *Red* locus are not certain, one intriguing possibility is that the dominant *R* allele (which produces red-cheeked birds) is maintained by negative frequency-dependent selection. Red-cheeked birds are behaviourally dominant (Pryke and Griffith 2006; Pryke 2007), and may out-compete black-cheeked birds for high quality nest holes (Brazill-Boast et al. 2013). However, as the frequency of Reds increases, they suffer fitness consequences because they provide worse parental care than black-cheeked birds in more competitive environments (Pryke and Griffith 2009c), and theoretical results suggest that, in combination with the assortative mating in this species, this is enough to maintain the a protected polymorphism (Kokko et al. 2014). However, without knowing the differences in fitness between phenotypes in the real world, it is impossible to say whether polymorphism at the *Red* locus is maintained in this way, as opposed to it being maintained by either heterozygote disadvantage, or heterozygote advantage.

One important goal of Chapter 4 was to determine whether selection for preferred codons is ongoing in *Drosophila*. Previous work in *D. melanogaster* (and to a lesser extent in *D. simulans*) had shown that the strength of selection for preferred codons has decreased since the common ancestor of *D. melanogaster* and *D. simulans* in both species, and may or may not currently be zero in *D. melanogaster* (Kliman 1999; McVean and Vieira 2001; Akashi et al. 2006). As mentioned above, the analyses in *D. melanogaster* were complicated by its non-equilibrium base composition, and much of the previous work was carried out on datasets with limited numbers of loci. By using 1) large polymorphism datasets from both species and 2) methods which account for non-equilibrium evolution, I showed evidence for ongoing selection for preferred codons in

both species, albeit to a lesser extent in *D. melanogaster*. I was also able to show that the selection favouring GC in *D. simulans* has remained relatively constant over the medium term, compared to *D. melanogaster*, which has undergone a reduction in the strength of selection favouring GC compared within the last $\sim 4N_e$ generations (Figure 4 in Chapter 4). This was achieved by using two complementary methodologies, one of which uses only polymorphic sites to estimate γ, and the other of which uses both polymorphic and fixed sites, and so is sensitive to signals further back in evolutionary time. This chapter serves to highlight the power of using between-species comparisons of evolutionary processes, as well as methods that are sensitive to these processes on different timescales.

Statistical issues

Some statistical concerns presented themselves during the course of this PhD. In Chapter 2, for instance, I showed that signatures of selection might be masked by the way in which information from multiple SNPs is combined when calculating F_{ST}. When taking the arithmetic mean of F_{ST} calculated per SNP (and thus weighting all SNPs equally; Eq (5) in Chapter 2), the action of purifying selection, which suppresses minor allele frequency at selected sites, is more apparent (Table 1 in Chapter 2). This signal is missed when information from SNPs is combined before taking the ratio of the between- and within-population components of F_{ST} (Eq (6) in Chapter 2), although this method may be more suitable when an overall measure of genetic differentiation is required. In Chapter 3, the non-random sampling at the *Red* locus posed a problem for the statistical power of our tests of the neutrality. For instance, the value of F_{ST} close to the focal site is expected to be nearly maximal under neutrality and the sampling scheme alone (Figure 4A in Chapter 4), regardless of whether or not selection has acted, and so increasing the probability of type II error (failing to correctly reject the null hypothesis). In Chapter 4, the parameter-richness of the models we used to i) infer ancestral states and ii) estimate γ, necessitated the combination of information from very many SNPs in order to avoid overfitting. Because of the dearth of diversity in our *D. melanogaster* dataset compared to our *D. simulans* dataset, we had to combine information from multiple bins to ensure against our results being affected by a lack of statistical power (Supplementary Table S3 in Chapter 4).

Final remarks

This thesis has used a mix of publicly available data, and those generated by collaborators; representing both large (genome-scale; Chapters 2 & 4) and smaller (large locus of interest and accompanying sex-linked reference loci; Chapter 3) datasets. What these chapters have in common is that for each I have used polymorphism and divergence data at different classes of site to try to understand the nature and extent of fundamental evolutionary forces, including demography, mutation and selection, that are acting in natural populations. All of these forces bear on our ability to investigate what is of interest in the natural world, either directly or indirectly. Sites that have traditionally been assumed to be of less interest to researchers, because they were thought not to influence phenotypes, are increasingly being shown to have functional importance. In Chapter 3, I showed that balancing selection is likely operating in an intergenic region on the Z chromosome in Gouldian finches. In Chapter 4, I showed evidence for ongoing selection at synonymous sites in *Drosphila melanogaster* and *D. simulans*. These results have implications for the detection of selection elsewhere in the genome, as well as inferences about demography and/or other evolutionary processes. The possibility of selection at so many extra sites may also bear on the magnitude of genetic load experienced by populations.

References

Aguade M, Miyashita N, Langley CH. 1989. Reduced Variation in the yellow-achaete-scute Region in Natural Populations of *Drosophila melanogaster*. Genetics 122:607–615.

Akashi H, Goel P, John A. 2007. Ancestral inference and the study of codon bias evolution: implications for molecular evolutionary analyses of the *Drosophila melanogaster* subgroup. PLoS One 2:e1065.

Akashi H, Ko W-Y, Piao S, John A, Goel P, Lin C-F, Vitins AP. 2006. Molecular evolution in the *Drosophila melanogaster* species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. Genetics 172:1711–1726.

Akashi H, Schaeffer SW. 1997. Natural Selection and the Frequency Distributions of Silent DNA Polymorphism in *Drosophila*. Genetics 146:295–307.

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics 139:1067–1076.

Akashi H. 1996. Molecular Evolution Between *Drosophila melanogaster* and *D. simulans* Reduced Codon Bias, Faster Rates of Amino Acid Substitution, and Larger Proteins in D. melanogaster. Genetics 144:1297–1307.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 19:711–722.

Andersson GE, Sharp PM. 1996. Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology 142:915–925.

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

Arguello JR, Zhang Y, Kado T, Fan C, Zhao R, Innan H, Wang W, Long M. 2010. Recombination yet inefficient selection along the *Drosophila melanogaster* subgroup's fourth chromosome. Mol. Biol. Evol. 27:848–861.

Auton A, McVean G. 2007. Recombination rate estimation in the presence of hotspots. Genome Res. 17:1219–1227.

Backström N, Zhang Q, Edwards S V. 2013. Evidence from a house finch (*Haemorhous mexicanus*) spleen transcriptome for adaptive evolution and biased gene conversion in passerine birds. Mol. Biol. Evol. 30:1046–1050.

Bakker EG, Toomajian C, Kreitman M, Bergelson J. 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. Plant Cell 18:1803–1818.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. Nat. Genet. 40:340–345.

Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. 13:969–980.

Beaumont MA, Nichols RA. 1996. Evaluating Loci for Use in the Genetic Analysis of Population Structure. Proc. R. Soc. B Biol. Sci. 263:1619–1626.

Beaumont MA. 2005. Adaptation and speciation: what can F(st) tell us? Trends Ecol. Evol. 20:435–440.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356:519–520.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLOS Biol. 5:e310.

Bergland AO, Tobler R, González J, Schmidt P, Petrov D. 2016. Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. Mol. Ecol. 25:1157–1174.

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

Betancourt AJ, Welch JJ, Charlesworth B. 2009. Reduced effectiveness of selection caused by a lack of recombination. Curr. Biol. 19:655–660.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting FST: The impact of rare variants. Genome Res. 23:1514–1521.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol. Biol. Evol. 21:1350–1360.

Bierne N, Eyre-Walker A. 2006. Variation in synonymous codon use and DNA polymorphism within the *Drosophila* genome. J. Evol. Biol. 19:1–11.

Bierne N, Roze D, Welch JJ. 2013. Pervasive selection or is it…? why are FST outliers sometimes so frequent? Mol. Ecol. 22:2061–2064.

Bolívar P, Mugal CF, Nater A, Ellegren H. 2015. Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill-Robertson interference, in an avian system. Mol. Biol. Evol. 33:216–227.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLOS Genet. 4:e1000083.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The Hitchhiking Effect on the Site Frequency Spectrum of DNA Polymorphisms. Genetics 140:783–796.

Brazill-Boast J, Griffith SC, Pryke SR. 2013. Morph-dependent resource acquisition and fitness in a polymorphic bird. Evol. Ecol. 27:1189–1198.

Brush AH, Seifreid H. 1968. Pigmentation and feather structure in genetic variants of the Gouldian finch, *Poephila gouldiae*. Auk 85:416–430.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157.

Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional Selection and the Site-Frequency Spectrum. Genetics 159:1779–1788.

Campo D, Lehmann K, Fjeldsted C, Souaiaia T, Kao J, Nuzhdin S V. 2013. Whole-genome sequencing of two North American *Drosophila melanogaster* populations

reveals genetic differentiation and positive selection. Mol. Ecol. 22:5084–5097.

Campos J, Charlesworth B, Haddrill P. 2012. Molecular evolution in nonrecombining regions of the *Drosophila melanogaster* genome. Genome Biol. Evol. 4:278–288.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol. Biol. Evol. 31:1010–1028.

Campos JL, Zeng K, Parker DJ, Charlesworth B, Haddrill PR. 2013. Codon usage bias and effective population sizes on the X chromosome versus the autosomes in *Drosophila melanogaster*. Mol. Biol. Evol. 30:811–823.

Caracristi G, Schlötterer C. 2003. Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. Mol. Biol. Evol. 20:792–799.

Carlini DB, Chen Y, Stephan W. 2001. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. Genetics 159:623–633.

Carvalho AB, Clark AG. 1999. Intron size and natural selection. Nature 401:344.

Casillas S, Barbadilla A, Bergman CM. 2007. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. Mol. Biol. Evol. 24:2222–2234.

Chamary J V, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. 7:98–108.

Chan AH, Jenkins PA, Song YS. 2012. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. PLoS Genet. 8:e1003090.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Structural approaches to sequence evolution. Springer. p. 207–232.

Charlesworth B, Charlesworth D, Barton NH. 2003. The Effects of Genetic and Geographic Structure on Neutral Variation. Annu. Rev. Ecol. Evol. Syst. 34:99–125.

Charlesworth B, Charlesworth D. 2010. Elements of evolutionary genetics. Greenwood Village: Roberts and Company Publishers

Charlesworth B, Morgan MT, Charlesworth D. 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics 134:1289–1303.

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. 70:155–174.

Charlesworth B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. 15:538–543.

Charlesworth B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. 77:153–166.

Charlesworth B. 2012a. The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila* X chromosome. Genetics 191:233–246.

Charlesworth B. 2012b. The effects of deleterious mutations on evolution at linked sites. Genetics 190:5–22.

Charlesworth B. 2013. Stabilizing selection, purifying selection, and mutational bias in finite populations. Genetics 194:955–971.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2:e64.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. Mol. Biol. Evol. 23:1348–1356.

Clarke B. 1970. Darwinian evolution of proteins. Science 168:1009–1011.

Clemente F, Vogl C. 2012a. Evidence for complex selection on four-fold degenerate sites in *Drosophila melanogaster*. J. Evol. Biol. 25:2582–2595.

Clemente F, Vogl C. 2012b. Unconstrained evolution in short introns? - an analysis of genome-wide polymorphism and divergence data from *Drosophila*. J. Evol. Biol. 25:1975–1990.

Collins TM, Wimberger PH, Naylor GJP. 1994. Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. Syst. Biol. 43:482–496.

Comeron JM, Kreitman M. 2000. The Correlation Between Intron Length and Recombination in *Drosophila*: Dynamic Equilibrium Between Mutational and Selective Forces. Genetics 156:1175–1190.

Comeron JM, Ratnappan R, Bailin S. 2012. The Many Landscapes of Recombination in *Drosophila melanogaster*. PLOS Genet. 8:e1002905.

Comeron JM, Williford A, Kliman RM. 2008. The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity 100:19–31.

Coop G, Griffiths RC. 2004. Ancestral inference on gene trees under selection. Theor. Popul. Biol. 66:219–232.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. PLOS Biol. 13:e1002112.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23:3133–3157.

Cutter AD, Charlesworth B. 2006. Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. Curr. Biol. 16:2053–2057.

Cutter AD, Choi JY. 2010. Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. Genome Res. 20:1103–1111.

Cutter AD, Moses AM. 2011. Polymorphism, divergence, and the role of recombination in *Saccharomyces cerevisiae* genome evolution. Mol. Biol. Evol. 28:1745–1754.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. 14:262–274.

Daborn P, Yen J, Bogwitz M, G LG, Feil E, Jeffers S, Tijet N, Perry T, Heckel D,

Batterham P, et al. 2002. A Single P450 Allele Associated with Insecticide Resistance in *Drosophila*. Science 297:2253–2256.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

David J, Capy P. 1988. Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. 4:106–111.

Dean AM, Thornton JW. 2007. Mechanistic approaches to the study of evolution: the functional synthesis. Nat. Rev. Genet. 8:675–688.

Dean MD, Ballard JWO. 2004. Linking phylogenetics with population genetics to reconstruct the geographic origin of a species. Mol. Phylogenet. Evol. 32:998–1009.

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.

Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S. 2013. Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. Genetics 193:291–301.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. 10:285–311.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12:640–649.

Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. Genome Res. 24:885–895.

Evans BJ, Zeng K, Esselstyn JA, Charlesworth B, Melnick DJ. 2014. Reduced representation genome sequencing suggests low diversity on the sex chromosomes

of Tonkean macaque monkeys. Mol. Biol. Evol. 31:2425–2440.

Excoffier L, Hofer T, Foll M. 2009. Detecting loci under selection in a hierarchically structured population. Heredity 103:285–298.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26:2097–2108.

Eyre-Walker A, Keightley PDP. 2007. The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8:610–618.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173:891–900.

Eyre-Walker A. 1998. Problems with Parsimony in Sequences of Biased Base Composition. J. Mol. Evol. 47:686–690.

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

Fabian DK, Kapun M, Nolte V, Kofler R, Schmidt PS, Schlötterer C, Flatt T. 2012. Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. Mol. Ecol. 21:4748–4769.

Fay JC, Wyckoff GJ, Wu C-I. 2001. Positive and Negative Selection on the Human Genome. Genetics 158:1227–1234.

Feder ME, Mitchell-Olds T. 2003. Evolutionary and ecological functional genomics. Nat. Rev. Genet. 4:651–657.

Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. Gene 463:18–20.

Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics

180:977–993.

Francino MP, Ochman H. 2001. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. Mol. Biol. Evol. 18:1147–1150.

Franklin DC, Dostine PL. 2000. Short Communication: A Note on the Frequency and Genetics of Head Colour Morphs in the Gouldian Finch. Emu 100:236–239.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220.

Fuller ZLZ, Haynes GGD, Zhu D, Batterton M, Chao H, Dugan S, Javaid M, Jayaseelan JC, Lee S, Li M, et al. 2014. Evidence for stabilizing selection on codon usage in chromosomal rearrangements of *Drosophila pseudoobscura*. G3 4:2433–2449.

Galtier N, Bazin E, Bierne N. 2006. GC-biased segregation of noncoding polymorphisms in *Drosophila*. Genetics 172:221–228.

Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. Genome Res. 25:1215–1228.

Griffiths RC. 2003. The frequency spectrum of a mutation, and its age, in a general diffusion model. Theor. Popul. Biol. 64:241–251.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. Genome Biol. 6:R67.

Haddrill PR, Charlesworth B. 2008. Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. Biol. Lett. 4:438–441.

Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. Genome Biol. 8:R18.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. 15:790–799.

Haddrill PR, Zeng K, Charlesworth B. 2011. Determinants of synonymous and nonsynonymous variability in three species of *Drosophila*. Mol. Biol. Evol. 28:1731–1743.

Hahn MWM. 2008. Toward a selection theory of molecular evolution. Evolution 62:255–265.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res. 16:875–884.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. PLoS Genet. 6:e1000825.

Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol. Biol. Evol. 24:2196–2202.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu. Rev. Genet. 42:287–299.

Hey J, Machado CA. 2003. The study of structured populations--new hope for a difficult and divided science. Nat. Rev. Genet. 4:535–543.

Hill W, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet. Res. 8:269–294.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat. Rev. Genet. 10:639–650.

Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res. 23:89–98.

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. 1994. Evidence for Positive Selection in the Superoxide Dismutase (Sod) Region of *Drosophila melanogaster*. Genetics 136:1329–1340.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164.

Hudson RR, Kreitman M, Aguade M. 1987. A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics 116:153–159.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of Levels of Gene Flow From DNA Sequence Data. Genetics 132:583–589.

Hudson RR. 1990. Gene Genealogies and the Coalescent. Oxford Surv. Evol. Biol. 7:1–44.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W. 2007. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. Genetics 177:469–480.

Innan H, Stephan W. 2003. Distinguishing the hitchhiking and background selection models. Genetics 165:2307–2312.

Jackson BC, Campos JL, Zeng K. 2015. The effects of purifying selection on patterns of genetic differentiation between *Drosophila melanogaster* populations. Heredity 114:163–174.

Jakobsson M, Edge MD, Rosenberg NA. 2013. The Relationship Between FST and the Frequency of the Most Frequent Allele. Genetics 193:515–528.

Johnston SE, Gratten J, Berenos C, Pilkington JG, Clutton-Brock TH, Pemberton JM, Slate J. 2013. Life history trade-offs at a single locus maintain sexually selected genetic variation. Nature 502:93–95.

Kaiser VB, Charlesworth B. 2009. The effects of deleterious mutations on evolution in non-recombining genomes. Trends Genet. 25:9–12.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation

accumulation lines. Genome Res. 19:1195–1201.

Keinan A, Mullikin JC, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat. Genet. 39:1251–1255.

Keinan A, Mullikin JC, Patterson N, Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat. Genet. 41:66–70.

Kern AD, Begun DJ. 2005. Patterns of polymorphism and divergence from noncoding sequences of *Drosophila melanogaster* and *D. simulans*: evidence for nonequilibrium processes. Mol. Biol. Evol. 22:51–62.

Kim K-W, Griffith SC, Burke T. 2016. Linkage mapping of a polymorphic plumage locus associated with intermorph incompatibility in the Gouldian finch (*Erythrura gouldiae*). Heredity 116:409–416.

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

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge, UK: Cambridge University Press

Kliman RM. 1999. Recent Selection on Synonymous Codon Usage in *Drosophila*. J. Mol. Evol. 49:343–351.

Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6:e15925.

Kokko H, Griffith SC, Pryke SR. 2014. The hawk-dove game in a sexually reproducing species explains a colourful polymorphism of an endangered bird. Proc. Biol. Sci. 281:20141794.

Kolaczkowski B, Kern AD, Holloway AK, Begun DJ. 2011. Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. Genetics 187:245–260.

Kreitman M, Hudson RR. 1991. Inferring the Evolutionary Histories of the Adh and

Adh-dup Loci in *Drosophila melanogaster* From Patterns of Polymorphism and Divergence. Genetics 127:565–582.

Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192:533–598.

Langley SA, Karpen GH, Langley CH. 2014. Nucleosomes Shape DNA Polymorphism and Divergence. PLoS Genet. 10:e1004457.

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. PLoS Genet. 9:e1003527.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2:e166.

Li W-H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24:337–345.

Li W-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36:96–99.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482.

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N, et al. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLOS Genet. 7:e1002326.

Marais G, Charlesworth B, Wright SI. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol. 5:R45.

Maruki T, Kumar S, Kim Y. 2012. Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms. Mol. Biol. Evol. 29:3617–3623.

Maside X, Lee AW, Charlesworth B. 2004. Selection on Codon Usage in *Drosophila americana*. Curr. Biol. 14:150–154.

Matsumoto T, Akashi H, Yang Z. 2015. Evaluation of Ancestral Sequence Reconstruction Methods to Infer Nonstationary Patterns of Nucleotide Substitution. Genetics 200:873–890.

Matsumoto T, John A, Baeza-Centurion P, Li B, Akashi H. 2016. Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. Mol. Biol. Evol. Online Ear.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23:23–35.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652–654.

McDowall M, Edwards NM, Jahoda CAB, Hynd PI. 2008. The role of activins and follistatins in skin and hair follicle development and function. Cytokine Growth Factor Rev. 19:415–426.

McVean GAT, Charlesworth B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. 74:145–158.

McVean GAT, Charlesworth B. 2000. The Effects of Hill-Robertson Interference Between Weakly Selected Mutations on Patterns of Molecular Evolution and Variation. Genetics 155:929–944.

McVean GAT, Vieira J. 2001. Inferring Parameters of Mutation, Selection and Demography From Patterns of Synonymous Site Evolution in *Drosophila*. Genetics 157:245–257.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. Proc. Natl. Acad. Sci. U. S. A. 110:8615–8620.

Mitchell-Olds T, Willis JH, Goldstein DB. 2007. Which evolutionary processes

influence natural genetic variation for phenotypic traits? Nat. Rev. Genet. 8:845–856.

Moriyama EN, Hartl DL. 1993. Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics 134:847–858.

Morton BR. 1998. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J. Mol. Evol. 46:449–459.

Myers SR, Griffiths RC. 2003. Bounds on the Minimum Number of Recombination Events in a Sample History. Genetics 163:375–394.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench-Constant RH, Blaxter ML, et al. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 367:343–353.

Nagylaki T. 1998. Fixation Indices in Subdivided Populations. Genetics 148:1325–1332.

Nei M, Miller JC. 1990. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics 125:873–879.

Nei M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. U. S. A. 70:3321–3323.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu S-A, Fraser D, et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337:100–104.

Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. Mol. Biol. Evol. 24:228–235.

Noor M a F, Bennett SM. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity 103:439–444.

Nordborg M, Charlesworth B, Charlesworth D. 1996. Increased Levels of

Polymorphism Surrounding Selectively Maintained Sites in Highly Selfing Species. Proc. R. Soc. B Biol. Sci. 263:1033–1039.

Nordborg M. 1997. Structured Coalescent Processes on Different Time Scales. Genetics 146:1501–1514.

O'Fallon BD, Seger J, Adler FR. 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. Mol. Biol. Evol. 27:1162–1172.

Ometto L, Stephan W, De Lorenzo D. 2005. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. Genetics 169:1521–1527.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat. Rev. Genet. 7:337–348.

Pannell JR. 2003. Coalescence in a Metapopulation with Recurrent Local Extinction and Recolonization. Evolution 57:949–961.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.

Paradis E. 2010. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26:419–420.

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

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. Mol. Biol. Evol. 27:1226–1234.

Parsch J, Tanda S, Stephan W. 1997. Site-directed mutations reveal long-range compensatory interactions in the Adh gene of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. 94:928–933.

Parsch J. 2003. Selective Constraints on Intron Evolution in *Drosophila*. Genetics 165:1843–1851.

Peden J. 1999. Analysis of codon usage. PhD Thesis. University of Nottingham, UK

Poh Y-P, Ting C-T, Fu H-W, Langley CH, Begun DJ. 2012. Population Genomic Analysis of Base Composition Evolution in *Drosophila melanogaster*. Genome Biol. Evol. 4:1245–1255.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ, et al. 2012. Population Genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture. PLoS Genet. 8:e1003080.

Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. Evolution 61:3001–3006.

Pool JE, Nielsen R. 2008. The impact of founder events on chromosomal variability in multiply mating species. Mol. Biol. Evol. 25:1728–1736.

Pryke SR, Griffith SC. 2006. Red dominates black: agonistic signalling among head morphs in the colour polymorphic Gouldian finch. Proc. Biol. Sci. 273:949–957.

Pryke SR, Griffith SC. 2007. The relative role of male vs. female mate choice in maintaining assortative pairing among discrete colour morphs. J. Evol. Biol. 20:1512–1521.

Pryke SR, Griffith SC. 2009a. Genetic incompatibility drives sex allocation and maternal investment in a polymorphic finch. Science 323:1605–1607.

Pryke SR, Griffith SC. 2009b. Postzygotic genetic incompatibility between sympatric color morphs. Evolution 63:793–798.

Pryke SR, Griffith SC. 2009c. Socially Mediated Trade-Offs between Aggression and Parental Effort in Competing Color Morphs. Am. Nat. 174:455–464.

Pryke SR. 2007. Fiery red heads: female dominance among head color morphs in the Gouldian finch. Behav. Ecol. 18:621–627.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org/ .

Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA:

contrasts among genes from *Drosophila*, mice, and humans. Mol. Biol. Evol. 13:735–748.

Reed FA, Akey JM, Aquadro CF. 2005. Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. Genome Res. 15:1211–1221.

Robinson MC, Stone EA, Singh ND. 2014. Population genomic analysis reveals no evidence for GC-biased gene conversion in *Drosophila melanogaster*. Mol. Biol. Evol. 31:425–433.

Roesti M, Hendry AP, Salzburger W, Berner D. 2012. Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. Mol. Ecol. 21:2852–2862.

Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. 2014. Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. Mol. Biol. Evol. 31:1750–1766.

Roy S, Ernst J, Kharchenko P V, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330:1787–1797.

Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. 2011. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. PLoS Genet. 7:e1001302.

Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW. 2005. The extent of linkage disequilibrium caused by selection on G6PD in humans. Genetics 171:1219–1229.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics 132:1161–1176.

Schaeffer S. 2002. Molecular population genetics of sequence length diversity in the Adh region of *Drosophila pseudoobscura*. Genet. Res. 80:163–175.

Schmidt JM, Good RT, Appleton B, Sherrard J, Raymant GC, Bogwitz MR, Martin J, Daborn PJ, Goddard ME, Batterham P, et al. 2010. Copy number variation and

transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. PLoS Genet. 6:e1000998.

Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427–1437.

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics 194:937–954.

Seger J, Smith WA, Perry JJ, Hunn J, Kaliszewska ZA, Sala L La, Pozzi L, Rowntree VJ, Adler FR. 2010. Gene genealogies strongly distorted by weakly interfering mutations in constant environments. Genetics 184:529–545.

Seidel HS, Rockman M V, Kruglyak L. 2008. Widespread genetic incompatibility in *C. elegans* maintained by balancing selection. Science 319:589–594.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? PLOS Genet. 5:e1000495.

Sharp PM, Li W-H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24:28–38.

Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125–5143.

Shields D, Sharp P, Higgins D, Wright F. 1988. “Silent” sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5:704–716.

Singh ND, Arndt PF, Clark AG, Aquadro CF. 2009. Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila*. Mol. Biol. Evol. 26:1591–1605.

Singh ND, Davis JC, Petrov DA. 2005. Codon bias and noncoding GC content correlate negatively with recombination rate on the *Drosophila* X chromosome. J. Mol. Evol. 61:315–324.

Singh ND, Macpherson JM, Jensen JD, Petrov DA. 2007. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster*. BMC Evol. Biol. 7:202.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol. Biol. Evol. 27:1813–1821.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature 415:1022–1024.

Southern HN. 1945. Polymorphism in *Poephila gouldiae* gould. J. Genet. 47:51–57.

Spencer CCA, Coop G. 2004. SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics 20:3673–3675.

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J. 2010. Adaptation genomics: the next generation. Trends Ecol. Evol. 25:705–712.

Stenøien HK. 2005. Adaptive basis of codon usage in the haploid moss *Physcomitrella patens*. Heredity 94:87–93.

Stephan W, Langley CH. 1989. Molecular Genetic Variation in the Centromeric Region of the X Chromosome in Three *Drosophila ananassae* Populations. I. Contrasts Between the vermilion and forked Loci. Genetics 121:89–99.

Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98:65–68.

Stevison LS, Noor MAF. 2010. Genetic and evolutionary correlates of fine-scale recombination rate variation in *Drosophila persimilis*. J. Mol. Evol. 71:332–345.

Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. Mol. Biol. Evol. 28:63–70.

Stone EA. 2012. Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines. Genome Res. 22:966–974.

Tajima F. 1989a. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tajima F. 1989b. The Effect of Change in Population Size on DNA Polymorphism. Genetics 123:597–601.

Takano-Shimizu T. 1999. Local Recombination and Mutation Effects on Molecular Evolution in *Drosophila*. Genetics 153:1285–1296.

Takano-Shimizu T. 2001. Local Changes in GC/AT Substitution Biases and in Crossover Frequencies on *Drosophila* Chromosomes. Mol. Biol. Evol. 18:606–619.

Tavaré S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.

Teshima KM, Innan H. 2009. mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. BMC Bioinformatics 10:166.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172:1607–1619.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. 39:31–40.

True JR, Mercer JM, Laurie CC. 1996. Differences in Crossover Frequency and Distribution Among Three Sibling Species of *Drosophila*. Genetics 142:507–523.

Turner TL, Levine MT, Eckert ML, Begun DJ. 2008. Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. Genetics 179:455–473.

Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. BMC Evol. Biol. 7:226.

Vogl C, Bergman J. 2015. Inference of directional selection and mutation parameters

assuming equilibrium. Theor. Popul. Biol. 106:71–82.

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

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276.

Weber C, C C, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. Genome Biol. 15:1–16.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358–1370.

Weir BS, Hill WG. 2002. Estimating F-statistics. Annu. Rev. Genet. 36:721–750.

Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. Genetics 173:821–837.

Wright S. 1951. The genetical structure of populations. Ann. Eugen. 15:323–354.

Wu C-I. 2001. The genic view of the process of speciation. J. Evol. Biol. 14:851–865.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yukilevich R, Turner TL, Aoki F, Nuzhdin S V, True JR. 2010. Patterns and processes of genome-wide divergence between North American and African *Drosophila melanogaster*. Genetics 186:219–239.

Zeng K, Charlesworth B. 2009. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. Genetics 183:651–662.

Zeng K, Charlesworth B. 2010a. Studying Patterns of Recent Evolution at Synonymous Sites and Intronic Sites in *Drosophila melanogaster*. J. Mol. Evol. 70:116–128.

Zeng K, Charlesworth B. 2010b. The effects of demography and linkage on the estimation of selection and mutation parameters. Genetics 186:1411–1424.

Zeng K, Charlesworth B. 2011. The joint effects of background selection and genetic

recombination on local gene genealogies. Genetics 189:251–266.

Zeng K, Corcoran P. 2015. The Effects of Background and Interference Selection on Patterns of Genetic Variation in Subdivided Populations. Genetics 201:1539–1554.

Zeng K, Fu Y-X, Shi S, Wu C-I. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174:1431–1439.

Zeng K. 2010. A simple multiallele model and its application to identifying preferred-unpreferred codons using polymorphism data. Mol. Biol. Evol. 27:1327–1337.

Zeng K. 2012. The application of population genetics in the study of codon usage bias. In: Cannarozzi GM, Schneider A, editors. Codon evolution: mechanisms and models. Oxford University Press. p. 245–254.

Zeng K. 2013. A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity 110:363–371.

Zhang C, Wang J, Long M, Fan C. 2013. gKaKs: the pipeline for genome-level Ka/Ks calculation. Bioinformatics 29:645–646.