Characterising the impact of directional forces on genomic variation

by

Henry Juho Barton

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

The University of Sheffield

Faculty of Science

Department of Animal and Plant Sciences

October 2019

Abstract

Mutation generates new genetic material on which evolutionary processes, such as selection and genetic drift, can act. Consequently, understanding mutations, and the forces that determine their fate, is vital for understanding how the variety of life seen today evolved. Technological and methodological advances over the past decade have lead to a growing availability of whole genome sequencing and re-sequencing datasets, allowing existing questions in genome evolution to be addressed more broadly, and previously understudied topics to be addressed. Two such areas are the selective landscape of insertions and deletions (INDELs), and the impact of GC biased gene conversion (gBGC) on non-coding base composition. To date, INDELs have remained understudied relative to point mutations, despite being the second most common form of variation. In part, this is due to the challenges of calling and correctly orientating INDELs, and in part due to a lack of methods available to quantify the selective pressures acting on them. Work on gBGC has largely concentrated on its action in coding regions, often with a view of how it confounds inferences of selection in these regions. In this thesis I make use of a number of publicly available datasets to extend the present state of knowledge in these two areas. Chapter 2 introduces a new method for inferring the distribution of fitness effects for INDELs, demonstrating its accuracy through simulations and using it to characterise the selective pressures on INDELs in coding sequences in *Drosophila melanogaster*. Chapter 3 applies the method to a dataset of great tit (*Parus major*) genomes and extends it to the non-coding regions in this species. Finally, Chapter 4 addresses the role and strength of gBGC across the non-coding genome of the great tit and the zebra finch (*Taeniopygia guttata*), and how gBGC and INDELs have contributed to base composition in these species.

Statement of intellectual contribution

The studies within the data chapters of this thesis have benefited from a number of collaborators as outlined below. The main chapters (Chapters 2 to 4) are presented in the form of scientific papers. The authors for all the papers are myself and Kai Zeng, who supervised the research contained in the chapters.

In Chapter 2, Kai Zeng derived the maximum likelihood model for estimating the distribution of fitness effects (DFE) for insertions and deletions, I carried out the simulation analyses and the preparation and analysis of the *Drosophila* dataset. The text of this chapter is as published in Barton and Zeng (2018).

In Chapter 3, I prepared the INDEL dataset and conducted all the analyses. The text of this chapter is as published in Barton and Zeng (2019).

In Chapter 4, an initial pipeline for generating the 1 mega base orthologous window dataset was provided by Pádraic Corcoran. I implemented a modified and extended version of this pipeline and constructed the rest of the work flow and performed the analyses. This chapter is presented as a manuscript that will be submitted for publication largely in its current form.

The artwork on the chapter title pages is by Carly Lynsdale.

Acknowledgements

Firstly I would like to thank Kai Zeng for his supervision of this thesis and from whom I have learnt a lot. I hope I haven't stressed you out too much over the years!

Secondly, my co-supervisor Jon Slate, whose supervision of a summer research project nurtured the initial seeds of my bioinformatic research interests, and who tolerated my unannounced problem solving meetings. I maybe took your advice less than I should, but it was nonetheless immensely helpful.

To the members of the B2212 and B2211 offices, past and present, you made the PhD that much richer.

Ben Jackson, my academic big brother, for your mix of logistical help and listening to me being melodramatic, thank you.

Leeban Yusuf, it was a brief but bright overlap, I enjoyed our lunches and bioinformatic solidarity.

Toni Gossmann. We may have had our programming differences, and I'm sorry for corrupting your students' coding styles (I'm not), but my time in Sheffield would not have been the same without you. Thank you for your guidance during both my masters and PhD, the emergency coffees, the pub trips, my (honorary?) position in the Gossmann lab, and all the laughs along the way.

Joe Baxter thank you for cohabiting with me, providing me with ample opportunity to refine my washing up skills, the games of pool, the beer, the whisky and spending 4 years in the dark since the light fitting broke.

Carly Lynsdale thank you for your patience and company throughout the ups and downs of the PhD, for your boundless enthusiasm and for running away underwater with me to hide from it all. It was, and remains, a pleasure!

Finally Pádraic Corcoran for working the bioinformatic coal face together, in sometimes hazardous, buggy, conditions. For suggesting an unlimited supply of new python modules each promised to be better than the last. For providing constant advice for the first two and half years of the PhD, and maintaining life debugging since. And of course facilitating my developing caffeine dependency, it helped with the mid week pub trips. Without you I suspect this thesis would not exist.

Contents

List of Figures

List of Tables

To Helen, KBO

Chapter 1

Thesis Introduction

1.1 Introduction

Mutation is the primary process by which new genetic variation arises. It generates the raw material on which evolutionary processes, such as selection and genetic drift, can act. Consequently, understanding the process of mutation, the fate of new mutations and the forces that determine that fate, is vital for understanding how the variety of life seen today evolved. Genetic studies, along with most of human endeavours, are constrained by the technologies of their time. Historically this has confined population genetic studies to small datasets, representing only a small proportion of a genome, such as individual genes or groups of genes (e.g. Blake *et al.*, 1992; Hess *et al.*, 1994), or short marker regions like microsatellites (e.g. Primmer *et al.*, 1997). Furthermore, the generation of these datasets and of initial larger scale genetic datasets was constrained to a select few species, such as humans (International Human Genome Sequencing Consortium, 2001) and model organisms such as *Caenorhabditis elegans* (C. elegans Sequencing Consortium, 1998).

However, with the advent of next generation sequencing technology (NGS) making cheap, high throughput sequencing possible (Zhang *et al.*, 2011), sequencing projects have exploded, with initiatives such as the Genome 10K project aiming to sequence 10,000 vertebrate genomes (Genome 10K Community of Scientists, 2009), the i5K Initiative aiming to sequence 5000 arthropod genomes (i5K Consortium, 2013), the Bird 10K project aiming to sequence 10,500 bird genomes (Zhang, 2015), the first 45 of which are already published (Jarvis *et al.*, 2014; Zhang *et al.*, 2014), and the Sanger Institute recently announced plans to sequence the genomes of 66,000 species native to the UK (Sanger Communications Team, 2018). However, these efforts are largely focussing on breadth, rather than depth, attempting to compare sequences from across the tree of life, but at the expense of high quality genome assemblies and the availability of individual resequencing data for a given individual species. Whilst this new scale of data available is allowing many core questions to be addressed more broadly, much of this work is still focused on selection operating on the most commonly studied form of mutation, point mutations, both substitutions and single nucleotide polymorphisms (SNPs).

Two topics with the potential to be greatly advanced by the availability of high quality resequencing datasets are the impact of selection on insertion and deletion mutations (INDELs), which have largely been overshadowed by SNPs, and GC biased gene conversion, the neutral process that behaves like selection favouring GC alleles. Here, I outline the present state of knowledge for both INDEL mutations and their selective landscape, as well as the process of GC biased gene conversion in the context of its role in the evolution of base composition. I also discuss these processes in the context of avian genomes, an ideal study system for advancing our knowledge of these topics.

1.2 Insertions and Deletions

1.2.1 Mechanism and Characteristics

The study of mutation and genetic variation in general has traditionally had SNPs in the limelight, as demonstrated by the development of SNP chips, making SNP analysis possible on an industrial scale (Syvänen, 2005). This focus has persisted through the paradigm shift that was the advent of next generation sequencing technology (Zhang *et al.*, 2011) to the present. Consequently, other forms of polymorphism, whilst not overlooked, have received less attention, and are consequently less well understood. A prime example is insertions and deletions (INDELs). In humans, INDELs are the second most common form of mutation after SNPs, and thus constitute a significant amount of genetic variation (Montgomery *et al.*, 2013). However, INDEL investigations face a number of stumbling blocks, perhaps explaining their lower level of research attention. One such difficulty is the need for multispecies alignments in order to distinguish insertion from deletion (for example see: Kvikstad and Duret, 2014). INDELs often occur in repetitive regions of the genome (Ananda *et al.*, 2013; Montgomery *et al.*, 2013), yet species alignment algorithms are weakest when aligning such regions (Earl *et al.*, 2014). Yet, with work towards resolving such issues underway (Earl *et al.*, 2014), and with the broad and expanding availability of whole genomes (Genome 10K Community of Scientists, 2009; i5K Consortium, 2013; Zhang, 2015), there is the unprecedented opportunity for more in-depth analysis of INDELs.

A number of mechanisms have been proposed for the generation of INDELs, including polymerase slippage (Garcia-Diaz and Kunkel, 2006; Levinson and Gutman, 1987) and improper repair of double stranded breaks (DSBs) (Chu, 1997). Polymerase slippage occurs when DNA denatures during replication, enabling the replicated and template strands to become misaligned. The resolution of such an event can give rise to an insertion if the replicated sequence 'slips' resulting in previously replicated bases being 're-replicated', or a deletion if the template strand slips and un-replicated bases are omitted (Figure 1.1) (Garcia-Diaz and Kunkel, 2006; Levinson and Gutman, 1987). Polymerase slippage has been suggested as the dominant force behind INDEL generation. Montgomery *et al.* (2013) analysed the sequence context of INDELs in a sample of 179 human genomes, revealing that 48% of INDELs fell within homopolymer runs, predicted hotspots near repeat regions, or tandem repeats. Furthermore, they demonstrate that 75% of all identified INDELs can be characterised as 'local changes in copy count', that is the loss or gain of a tandem duplicate. This percentage ranged from 56% outside of repetitive regions up to 95% within dinucleotide tandem repeats. These patterns are consistent with polymerase slippage driving the majority of INDEL mutation. Montgomery *et al.* (2013) suggest the remaining 25% of INDELs are likely a result of improper repair of DSBs (see Chu, 1997, for review). This is supported by Drosophila mutants with reduced DSB repair capabilities demonstrating a correlation between DSB repair errors and large deletions (McVey *et al.*, 2004). Additionally, a small number of INDELs ($\sim$ 1.3% in humans) contain palindromic repeats, consistent with a brief switch, and subsequent recovery, of the template strand as a mechanism of formation for these INDELS (Montgomery *et al.*, 2013).

If polymerase slippage is the main mechanism driving INDEL generation, then it could explain the bias towards deletions over insertions observed in many organisms. This bias, calculated as the ratio of deletions to insertions (rDI) is reported from both polymorphism data (rDI$_{pol}$) and divergence data (rDI$_{div}$). Both measures of rDI are around 2 in humans (Kvikstad and Duret, 2014; Nam and Ellegren, 2012). In *Arabidopsis spp.* rDI$_{div}$ is 1.4 and rDI$_{pol}$ is as high as 8 (Hu *et al.*, 2011). In *Drosophila spp.* rDI$_{div}$ is 0.8 and 2.7 for *D. melanogaster* and *D. simulans* respectively, whilst rDI$_{pol}$ is 1.18 in

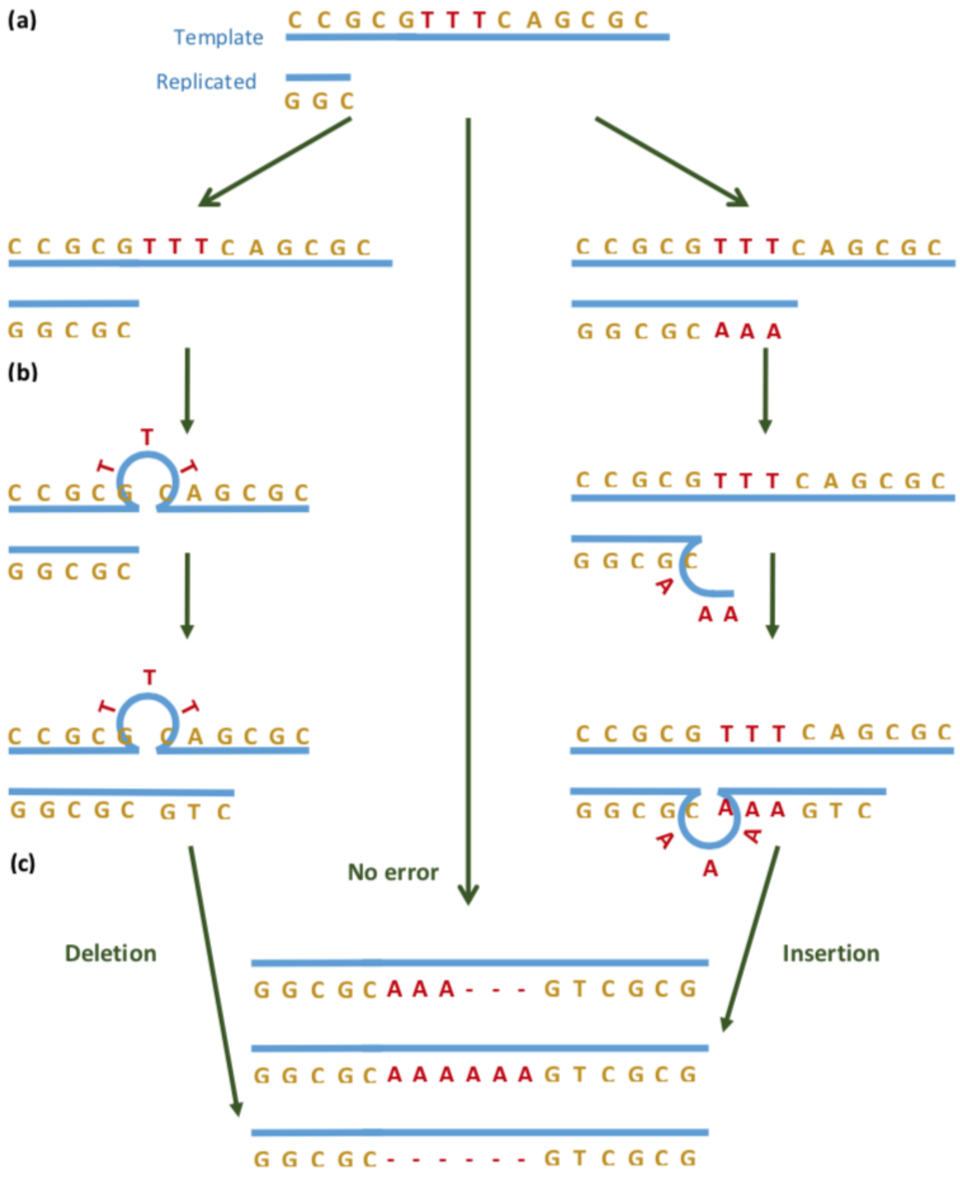

(a)
Template
C C G C G T T T C A G C G C
Replicated
G G C
C C G C G T T T C A G C G C
G G C G C
(b)
C C G C G T T T C A G C G C
G G C G C A A A
C C G C G T T T C A G C G C
G G C G C
G G C G C A A
C C G C G T T T C A G C G C
G G C G C G T C
G G C G C A A A G T C
(c)
No error
Deletion
Insertion
G G C G C A A A - - - G T C G C G
G G C G C A A A A A A G T C G C G
G G C G C - - - - - - G T C G C G

Figure 1.1: INDEL formation through polymerase slippage. During replication (a) template and replicated strands can denature, resulting in strand slippage at the polymerase and formation of loops (b), giving rise to a deletion or insertion if the template or replicated strand respectively slips (c).

D. melanogaster (Presgraves, 2006). As generation of deletions requires DNA to denature at a single position only, whereas insertions require a duplicated strand the length of the insertion to denature, deletions might be expected to occur more often by this mechanism (Petrov, 2002b). This deletion bias, and the magnitude of the rDI have been suggested to be the fundamental driving force behind the evolution of genome size, with higher ratios resulting in more rapid genome contraction (Petrov, 2002b). Petrov's

(2002b) mutational equilibrium hypothesis suggests that genome expansion is predominantly determined by the size of small deletions. This is based on two assumptions; firstly that selection quickly removes long deletions and secondly that selection prevents large increases in INDEL error rate. Thus, Petrov (2002b) proposes that genome retraction or expansion is determined by the balance between small deletion size and large insertions (required to make up for the low insertion rate). The model gains support from a correlation between genome size and deletion bias (Petrov, 2002b). However, it has also been the subject of contention. Gregory (2003, 2004) suggests that this correlation is primarily driven by *Drosophila spp.* and does not extend well to other organisms. Even if this is not the case, it has been proposed that deletion bias and the action of small deletions would bring about changes in genome size so slowly that it is not sufficient to explain genome size variation between species within known divergence times. Large insertions and deletions however could explain more rapid size changes (Gregory, 2003, 2004).

The rDI$_{div}$, however, may not solely be the product of biased mutation (i.e. under neutrality rDI$_{div}$ should equal rDI$_{pol}$, but this is often not the case), but due to variation in fixation probabilities of insertions and deletions. In *Arabidopsis spp.* deletions are observed to have elevated fixation rates (Hu *et al.*, 2011). However, in humans, fixation rate is generally higher for insertions, but with some non-repetitive regions showing evidence for both deletions experiencing increased fixation (Kvikstad and Duret, 2014). These observed differences in fixation probability between INDELs could be driven by either selection or biased gene conversion. Analysis of *Drosophila melanogaster* introns and intergenic sequences revealed a higher insertion fixation rate in introns than intergenic sequence. This is theorised to be due to selective constraint on intron length, with insertions selected for to prevent intron length being reduced below a minimum length (Ometto *et al.*, 2005). The role of this elevated fixation rate in driving intron length change is supported in a comparison of *D. melanogaster* and *Drosophila simulans* introns. The amount of deletions in each species is seen to be similar, but *D. melanogaster*, with the longer introns of the two species, has a significantly greater number of fixed insertions. However, this trend was most pronounced on the X chromosome, which has an elevated crossing-over rate, suggesting that insertion biased gene conversion (iBGC) may be a better explanation for elevated insertion fixation rates than selection (Presgraves, 2006). Conversely, intron length has also been seen to negatively correlate with recombination rate in *D. melanogaster* (Comeron and Kreitman, 2000).

More recently, iBGC has garnered support from an analysis of non-coding INDELs in humans, *D. melanogaster* and *Saccharomyces cerevisiae*. In agreement with previous work, these species were seen to have increased fixation probabilities for insertions. Yet, in an advancement on previous studies, the relationship of this fixation bias with recombination was investigated, with rDI_{div} for small (1-4bp) INDELs showing a negative correlation with recombination rate, whilst rDI_{pol} is nearly independent. Thus, fixation probability for small insertions is highest in highly recombining regions, a trend indicative of iBGC (Leushkin and Bazykin, 2013). Whilst both selection and iBGC have been suggested to be responsible for this trend, Kvikstad and Duret (2014) offer a third explanation. They show that incorrect polarisation of INDELs resulting from misidentification of derived and ancestral sequences could also explain the observed insertion fixation bias. This parallels the issue with evidence for GC biased gene conversion (gBGC, discussed later) (Hernandez *et al.*, 2007). Inaccurate polarisation of INDELs is further compounded by the possibility of INDEL hotspots, as suggested in humans, where multiple occurrences of INDELs at a single locus confuses polarisation (Kvikstad and Duret, 2014). That said, there is support from mutation accumulation experiments for a deletion bias in both *Drosophila spp.* (Keightley *et al.*, 2009) and *Arabidopsis spp.* (Ossowski *et al.*, 2010), where incorrect INDEL polarisation is unlikely, suggesting that the phenomenon is not purely artefactual.

As mentioned above, INDELs cluster in particular genetic regions, or hotspots. In analyses of a dataset of 179 human genomes, 40 to 48 percent of identified INDELs were seen to occur in repeat regions, even though these repeat regions comprise only three to four percent of the genome (Ananda *et al.*, 2013; Montgomery *et al.*, 2013). This is not surprising, as polymerase slippage is positively correlated with the length of identical repeat sequences (Klintschar and Wiegand, 2003). However, INDEL hotspots are not confined to repetitive regions. Work on chimpanzee (*Pan troglodytes*), orangutan (*Pongo abelii*), rhesus macaque (*Macaca mulatta*) and human genomes established that

14% of INDEL loci in non-repetitive regions are ‘complex’. That is, they have been subject to two or more INDEL events, and thus can be considered hotspots (Kvikstad and Duret, 2014). In addition, there is evidence that INDELs themselves may generate SNP hotspots. Tian *et al.* (2008), in a broad study comparing genomes of primates, rodents, rice and yeasts, showed that nucleotide diversity increases in regions neighbouring INDELs. They constructed a model based on the concept that if INDELs are mutagenic, they should be so only when they are in a heterozygous individual, so as to disrupt chromosomal pairing during recombination. The model predicts that there should be significantly more SNPs associated with the INDEL containing allele than the ancestral allele. Indeed, this is what the authors report in an analysis of 1027 INDELs in three yeast strains (Tian *et al.*, 2008).

1.2.2 Insertions and Deletions in Birds

INDEL research thus far has predominantly been focused on model organism such as *Drosophila* (Keightley *et al.*, 2009; Leushkin and Bazykin, 2013; Petrov, 2002b; Presgraves, 2006), *Arabidopsis* (Hu *et al.*, 2011; McVey *et al.*, 2004; Ossowski *et al.*, 2010) and chimpanzees, along with other primates, including humans (Ananda *et al.*, 2013; Kvikstad and Duret, 2014; Montgomery *et al.*, 2013). However, with the ongoing explosion in available bird genomes (Jarvis *et al.*, 2014; Zhang, 2015; Zhang *et al.*, 2014), it is now possible to investigate INDELs genome wide in multiple bird species. Birds are particularly suited for such investigation for two main reasons. Firstly, they have very stable genomes with conserved karyotypes (see van Oers *et al.*, 2014) and few repeat elements (Primmer *et al.*, 1997). Hence, it is interesting how INDEL variation evolves in a system where changes in genome length are not common and repeats are less frequent. Secondly, birds are noted for their highly variable recombination rates, with recombination being elevated on micro-chromosomes and towards telomeres, and reduced on macro-chromosomes and towards centromeres (Backström *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014). As such, birds present a good opportunity to understand the importance of proposed phenomena such as iBGC (Leushkin and Bazykin, 2013) and selection on INDELs to maintain a minimum intron length (Ometto *et al.*, 2005).

To date there are only a handful of published studies on INDELs in birds. Of these studies, those that investigate INDELs at the whole genome level are focussed on identifying INDELs in chicken, with a view of uncovering their role in commercially important traits and disease (Boschiero *et al.*, 2015; Yan *et al.*, 2014). These analyses report an rDI$_{pol}$ in chicken of around 1.3 (Boschiero *et al.*, 2015; Yan *et al.*, 2014), in agreement with earlier work reporting an rDI$_{pol}$ of 1.4 (Brandstrom and Ellegren, 2007), but not as high as the estimate of rDI$_{div}$ by Nam and Ellegren (2012) of 3.2. In addition, estimates of the deletion to insertion ratio in other avian species vary hugely, rDI$_{div}$ is estimated at 6.3 in pigeons and doves (Johnson, 2003), 3.5 in zebra finch (*Taeniopygia guttata*) (Nam and Ellegren, 2012) and with two studies of more than 40 species reporting mean rDI$_{div}$ estimates of 2.5 (Paśko *et al.*, 2011; Sundström *et al.*, 2003). Variation in deletion bias has also been reported between chromosomes, with estimates for rDI$_{div}$ on the Z and W chromosomes of 1.9 and 7.3 respectively (Sundström *et al.*, 2003). Thus, though there is much variation in the reported magnitudes of avian deletion bias, its occurrence is clear.

Beyond deletion bias, a number of studies have also reported intra-genomic variation in INDEL density in birds. INDEL density is reportedly lowest on micro-chromosomes (Boschiero *et al.*, 2015; Brandstrom and Ellegren, 2007; Yan *et al.*, 2014), with Brandstrom and Ellegren (2007) estimating INDEL density on macro-chromosomes to be 20% higher. However, they also report a strong correlation between INDEL density and SNP density, a finding at odds with that of Boschiero *et al.* (2015) who report both low INDEL density and high SNP density on micro-chromosomes. In addition, elevated mutation rates have been reported on micro-chromosomes (Axelsson *et al.*, 2005), in-keeping with the latter scenario. It has been suggested that the lower INDEL density on micro-chromosomes is due to elevated gene density on shorter chromosomes, making INDELs more likely to be strongly deleterious (Yan *et al.*, 2014). In support of this idea, the rate of insertions in chicken and zebra finch has been found to negatively correlate with gene density (Nam and Ellegren, 2012). However, the opposite relationship has been reported for the rate of small deletions, which correlates positively with gene density. In addition, both gene density and the small deletion rate are seen to correlate with recombination rate. It has therefore been suggested that there may be a causal relationship driving the small size of micro-chromosomes. Where, as a chromosome shrinks, recombination rate increases, in turn increasing the small deletion rate, resulting in further chromosome shrinking and so on (Nam and Ellegren, 2012). Such a mechanism, would lend support to the theory proposed by Petrov (2002b) to explain the evolution of genome size.

In addition to INDEL density being lower on micro-chromosomes (Boschiero *et al.*, 2015; Brandstrom and Ellegren, 2007; Yan *et al.*, 2014), INDEL density has been shown to be reduced on the Z chromosome (Brandstrom and Ellegren, 2007; Yan *et al.*, 2014). This reduction in INDEL density over autosomes has been estimated at 45% (Yan *et al.*, 2014). Furthermore, analysis of INDEL fixation rates on the avian sex chromosomes reveals the Z chromosomes INDEL fixation rate to be twice that of the W chromosome. Considering that males are homogametic, ZZ, this suggests a male INDEL bias in birds (Sundström *et al.*, 2003). Interestingly, in addition to having lower INDEL density than autosomes, and elevated INDEL rate relative to W, the Z chromosome also has a markedly lower deletion bias of 1.9 than the W chromosomes 7.3 (Sundström *et al.*, 2003). The W chromosomes history is marked by chromosome reduction following the prevention of recombination with Z (Sundström *et al.*, 2003). Therefore, the elevated deletion bias on W could be taken as indirect evidence for the action of iBGC, where the lack of recombination has stopped iBGC, allowing deletions to occur unchecked, driving W chromosome reduction. Some weak support can be derived for iBGC in birds from a reported positive correlation between GC content and INDEL density (Brandstrom and Ellegren, 2007), as GC content is also correlated with recombination rate, a relationship suggested to be due to gBGC (Bolívar *et al.*, 2016). However, there is also evidence against this, with a negative correlation between recombination and insertion fixation rate also reported in birds (Nam and Ellegren, 2012).

Finally, work on avian INDELs supports polymerase slippage as the main mechanism for INDEL generation. Chicken INDELs are reported to be A rich, consistent with increasing AT content lowering sequence melting temperature and increasing the chance of misalignment slippage (Brandstrom and Ellegren, 2007; Fryxell and Zuckerkandl, 2000). The findings discussed here clearly show many conflicting narratives in the existing avian INDEL literature. With agreement only on the existence of a deletion bias (Boschiero *et al.*, 2015; Brandstrom and Ellegren, 2007; Johnson, 2003; Nam and Ellegren, 2012; Paśko *et al.*, 2011; Sundström *et al.*, 2003; Yan *et al.*, 2014) and lower INDEL density on micro-chromosomes (Boschiero *et al.*, 2015; Brandstrom and Ellegren, 2007; Yan *et al.*, 2014). In addition, most of the published avian INDEL studies focus on domesticated species such as chicken (Boschiero *et al.*, 2015; Brandstrom and Ellegren, 2007; Nam and Ellegren, 2012; Yan *et al.*, 2014) and zebra finch (Nam and Ellegren, 2012), with little work on the selective pressures on INDELs in a wild system. However, the recent availability of whole genomes for many wild bird species (Jarvis *et al.*, 2014; Zhang *et al.*, 2014) has created an unprecedented opportunity to study the role of selection on INDELs in the evolution of avian genomes.

1.3 The Determinants of Genomic Base Composition

Most work to date has made use of point mutations to investigate the evolution of GC content within genomes. From this perspective changes in base composition are ultimately the result of biases in the fixation rate, if more G and C polymorphisms fix than As and Ts, then GC content will increase, and vice versa. Under neutrality the fixation rate is equal to the mutation rate, thus mutational biases translate into fixation biases and can alter GC content. However, if the sequence in question is not evolving truly neutrally, then fixation rates can be biased independently of mutation. In this section I will review mutation rate variation and fixation biasing processes as drivers of base composition evolution, with a focus on GC biased gene conversion.

1.3.1 Mutation Rate Variation

Since the 1960s it has been known that mutation rate is not constant within genomic sequences (Benzer, 1961), since then, much has been revealed. Mutation rate variation has been characterised within and between mammalian mitochondrial genomes, and is seen to evolve rapidly (Galtier, 2005). However, it was not until the advent of next generation sequencing (NGS) that the question of mutation rate variation could be addressed more comprehensibly (for review see Zhang *et al.*, 2011). NGS made it possible to sequence whole genomes for a previously unimaginable range of species. This explosion in obtainable whole genome data has allowed research to more thoroughly investigate the mechanisms and causes of mutation rate variation.

Fine scale genomic variation in mutation rates has been largely explained by a number of contextual explanations. Firstly, transitions, mutations between purines (A and G nucleotides) or between pyrimidines (T and C nucleotides) are twice as common as transversions (mutations between purines and pyrimidines), despite there being twice as many types of the latter than the former (Blake *et al.*, 1992; Keightley *et al.*, 2009; Ossowski *et al.*, 2010; Ségurel *et al.*, 2014). Secondly, strong bases, that is C or G, are twice as likely to mutate to weak bases, T or A, than weak are to strong (Hodgkinson and Eyre-Walker, 2011; Hwang and Green, 2004; Ségurel *et al.*, 2014). Both these trends are seen across a wide range of organisms, including mammals (Blake *et al.*, 1992; Hwang and Green, 2004), *Drosophila melanogaster* (Keightley *et al.*, 2009) and *Arabidopsis thaliana* (Ossowski *et al.*, 2010). The trend of strong-to-weak mutation bias is largely driven by CpG sites. Methylated C readily mutates to T, resulting in an under-representation of CpG in the human genome (Bird, 1980). Interestingly, CpG mutability has a negative relationship with GC content, with methylated C stability declining with GC content, as DNA melting temperature is also reduced (Fryxell and Zuckerkandl, 2000). This has been suggested to result in an escalating cycle of GC decrease. Reductions in GC content lower the sequences melting temperature, promoting C to T mutations, which further reduces GC content. The reverse can occur when GC content experiences an increase, raising melting temperature and reducing C to T mutations (Fryxell and Zuckerkandl, 2000). Thus, when it is considered that CG sites have a 15-fold increase in mutation rate relative to other sites (Hodgkinson and Eyre-Walker, 2011), and that C to T mutation increases two fold when GC content is reduced by 10 percent (Fryxell and Zuckerkandl, 2000), this cycle can bring about significant variation in mutation rates within and between genomes. It is therefore unsurprising that CpG hypermutability is implicated in an ongoing reduction in GC content in mammalian genomes (Duret *et al.*, 2002), though it has more recently been suggested this may not hold across all mammals (Romiguier *et al.*, 2010). Additionally, mutation accumulation experiments in *Arabidopsis thaliana* show unmethylated GC sites also had elevated mutability, suggesting other mechanisms may be contributing to this trend (Ossowski *et al.*, 2010).

Whilst CpG sites are among the strongest contextual determinants of mutation rate, accounting for 19% of mutations despite making up less than 2% of sites in humans, there are a number of other important context dependent effects (Ségurel *et al.*, 2014). One such phenomenon is 'neighbour effects'; the impact of base content on the mutation rate of the adjacent base (see Hwang and Green, 2004). This has been studied by comparing nucleotide frequencies neighbouring neutral substitutions with expected base frequencies, and has yielded conflicting results. Early research on human pseudogenes suggested that neighbour effects were strongest when the $5'$ neighbour was A or the $3'$ neighbour was G, and weakest when the $3'$ neighbour was C, though conflicting results were obtained for the weakest $5'$ neighbour effect (Blake *et al.*, 1992; Hess *et al.*, 1994). Later genome wide analysis supports the relative strength of the neighbour effect of a $3'$ G (Zhao and Boerwinkle, 2002), as might be expected with the high mutability of methylated C at CpG sites (Bird, 1980), but disagrees with the effect of all other bases (Zhao and Boerwinkle, 2002). It may be that disagreement stems from the oversimplification of context effect scenarios. Hwang and Green (2004) advocate the separation of neighbour effects into 14 classes, based on the analysis of context effects in 19 mammal species. Whilst, GC content aside, agreement has yet to be reached on the importance of individual bases and base contexts, neighbour effects are undeniably influencing local substitution rates. Although A and T nucleotides, on average, have a reduced mutability relative to G and C (Hodgkinson and Eyre-Walker, 2011; Hwang and Green, 2004; Ségurel *et al.*, 2014), they are not beyond implication in mutation rate variation. Blake *et al.* (1992) observed an increased frequency of TA to CG transitions than expected in humans. Similarly, A to G transitions are seen 40 percent more often on the coding strand than the non-coding strand in humans (see Ségurel *et al.*, 2014). Thus overall mutation rates tend to be biased towards strong to weak mutations as driven by CpG mutability and act to reduce genomic GC content.

1.3.2 Fixation Rate Variation

Whilst under neutrality the fixation rate is equal to the mutation rate, large swathes of genomes do not behave neutrally, both in coding and non-coding regions. As a result, fixation rates can deviate from mutation rates under such conditions. Yet, generally selection is not expected to favour GC alleles over AT alleles, so should not unduly influence genomic base composition. However, there is an exception in the case of translational selection driving biases is synonymous codon usage (Sharp *et al.*, 1995) as well as with the neutral process of GC biased gene conversion (Chen *et al.*, 2007).

1.3.2.1 Codon Usage Bias and Translational Selection

One phenomenon that may cause biased fixation rates in coding regions is translational selection resulting in codon usage bias. The degeneracy of the genetic code results in synonymous codons, i.e. codons that differ in sequence but code for the same amino acid. The translational selection model proposes that preferred codons (those synonymous codons occurring at elevated frequency) are those for which tRNAs are most abundant, and thus those that can be translated most efficiently (Duret, 2002; Sharp *et al.*, 1995). Under the translational selection model, preferred codon usage should correlate with expression level, and indeed this is supported by evidence of codon bias increasing with gene expression in a number of model organisms such as *Drosophila spp.*, *Caenorhabditis elegans* and *Arabidopsis thaliana* (Bierne and Eyre-Walker, 2006; Duret and Mouchiroud, 1999). More recently, Galtier *et al.* (2018) used a dataset consisting of 30 species, each from a separate metazoan family, to identify a general preference for C over G terminating codons and T over A ending codons, when analysing GC conservative synonymous codon pairs. The authors suggest this preference for pyrimidines may be linked to them sharing a tRNA unlike purines. Whilst translational selection is likely an important component of codon usage bias and thus influences base composition, GC biased gene conversion is emerging as an increasingly important determinant of codon usage bias (Galtier *et al.*, 2018; Jackson *et al.*, 2017), as well as genome evolution in general (Chen *et al.*, 2007).

1.3.2.2 GC Biased Gene Conversion

GC biased gene conversion (gBGC), unlike codon usage bias, is not confined to coding regions. gBGC drives increases in GC allele frequency and occurs during recombination as follows. Firstly, recombination is initiated by a double stranded break (DSB) in one of a pair of homologous chromosomes (figure 1.2b). One strand of the broken chromosome invades its counterpart leading to the formation of a 'D loop' (figure 1.2c). At which point recombination and gene conversion can proceed in a number of directions (for more information see: Chen *et al.*, 2007; Duret and Galtier, 2009), here for simplicity I focus on the process that results in a crossover. DNA synthesis proceeds along the length of the D loop, resulting in a crossover, which leaves heteroduplex regions, where the two stands of a sequence are from different parent chromatids (figure 1.2d). If there are heterozygous sites within this heteroduplex, it results in miss-paired bases (outlined bases in figure 1.2d). The repair of these mismatches, gene conversion, favours GC alleles over AT alleles, hence 'GC biased' gene conversion, allowing these alleles to increase in frequency in heterozygotes in a selection mimicking manner (Chen *et al.*, 2007; Duret and Galtier, 2009; Galtier and Duret, 2007).

As gBGC operates like selection and occurs during recombination, both recombination rate and the effective population size (N_e) are determinants of its strength and impact. As recombination rate increases, so should the number of gene conversions. Whereas N_e modulates the strength of genetic drift, and thus how effectively gBGC can increase the frequency of GC alleles. Therefore, in areas of low N_e and areas of low recombination, the impact of gBGC should be negligible, and in areas where these parameters are high the effects of gBGC should be pronounced. To explore the role of gBGC in genomic evolution, studies largely contrast the rate of substitution for three categories of site. Firstly, substitutions from weak bases (A and T) to strong bases (G and C) (WS), which are elevated by gBGC. Secondly, substitutions in the opposite direction from strong to weak bases (SW), for which gBGC reduces the fixation probability. Finally weak to weak and strong to strong substitutions (WWSS), also known as GC conservative changes, which will not be impacted by gBGC. Comparison of the ratios of WS substitutions to SW substitutions across 19 mammals, showed higher ratios, indicative of gBGC, in

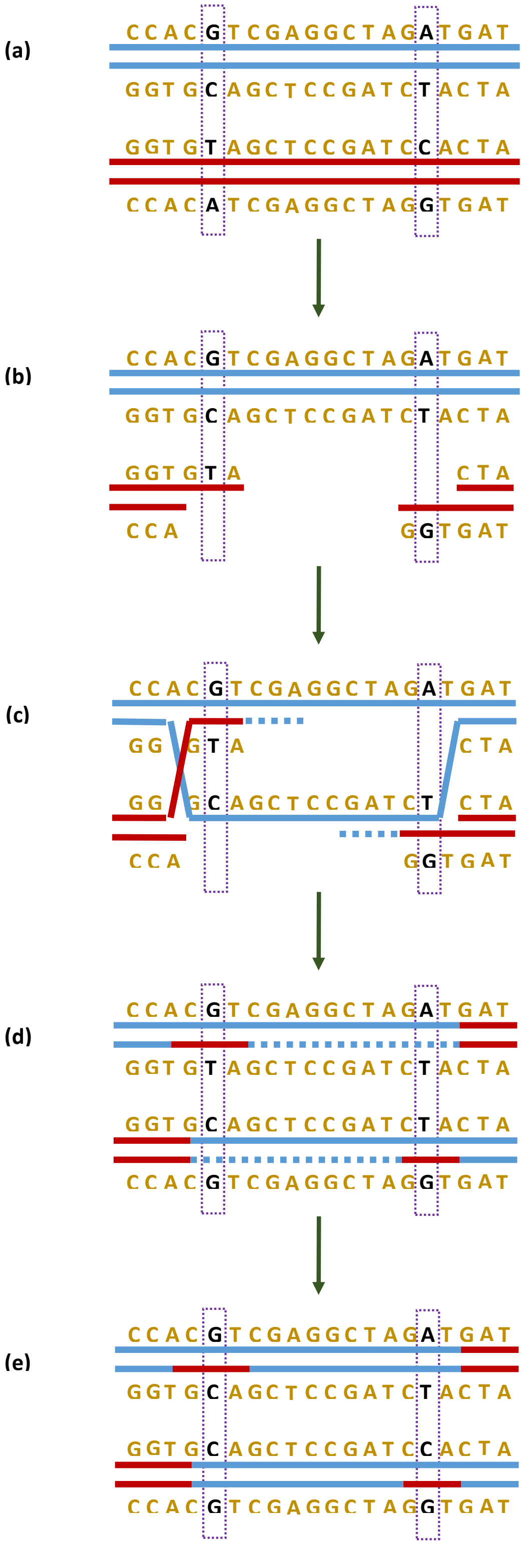


Figure 1.2: A simple overview of gene conversion. Recombination is initiated by a double stranded break (b), which is followed by the formation of a 'D loop' and strand invasion (c) which is resolved leading to a crossover (d), finally the heteroduplex mismatches (outline bases in black) are resolved (e).

species with higher N_e (Hwang and Green, 2004). WS substitution rates have also been shown to positively correlate with recombination rate in the fly catcher (*Ficedula albicollis*) (Bolívar *et al.*, 2016), and with GC content in *Drosophila melanogaster* and *Drosophila simulans* (Jackson *et al.*, 2017). Evidence for the action of gBGC can be seen through WS polymorphic SNPs segregating at higher frequency than other mutation types across a range of birds and mammals (Rousselle *et al.*, 2019). gBGC also provides an explanation for the relationship between recombination rate and GC content reported across a broad range of taxa (Bolívar *et al.*, 2016; Glémin *et al.*, 2015; Rousselle *et al.*, 2019; Wallberg *et al.*, 2015; Weber *et al.*, 2014).

An alternative, but less applied approach, is to make use of polymorphism datasets and estimate the population scaled strength of gene conversion B (where $B = 4N_eb$). This is analogous to estimating the population scaled selective coefficient ($\gamma = 4N_es$) for WS mutations. Long *et al.* (2018) recently calculated B across a wide range of species, spanning multiple taxa, yielding a range of B from 0.4 to 5. When considering previous estimates, humans are at the lower end of this range, with a mean B of 0.38 (Glémin *et al.*, 2015), *D. melanogaster* and *D. simulans* span the lower half of the range with B values ranging from 0.5 to 2.5 within their genomes (Jackson *et al.*, 2017) and the honey bee (*Apis mellifera*) sits at the top with a mean B of 5.7 (Wallberg *et al.*, 2015). Although, as indicated by the large range of intra-genomic B estimates in *Drosophila* (Jackson *et al.*, 2017), B will vary depending on genomic context, and so localised B values may fall greatly outside the range reported by Long *et al.* (2018), such as in humans where B in recombination hotspots is as high as 18 (Glémin *et al.*, 2015). In keeping with analyses of WS substitution patterns, B correlates with recombination rate (Glémin *et al.*, 2015; Wallberg *et al.*, 2015) GC content (Jackson *et al.*, 2017) and N_e (preprint: Borges *et al.*, 2018).

In addition to characterising how historical (substitution based methods) and contemporary (polymorphism based methods) gBGC proceeds, many studies also assess the long term impact of gBGC on base composition, by calculating the equilibrium GC content (GC$*$). GC$*$ is an estimate of the GC content if gBGC continues at the same strength over an indefinite branch length. GC$*$ can be estimated from either polymorphism or substitution based methods. Studies to date generally report strong correlations between GC$*$ and predictors of gene conversion, such as recombination rate (Muyle *et al.*, 2011; Singhal *et al.*, 2015), GC content (Gossmann *et al.*, 2018) and N_e (Weber *et al.*, 2014). Such findings show how elevated gBGC levels are driving GC content to higher proportions in some regions and species.

1.3.2.3 GC Biased Gene Conversion in Birds

As discussed previously, avian genomes are a particularly useful tool in addressing many population genetic questions in part due to their conserved synteny and karyotype (Hansson *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014; Zhang *et al.*, 2014), facilitating between species comparisons, and their variable recombination landscape (Backström *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014), making them particularly suited to studying recombination driven processes like gene conversion. Furthermore, it has been suggested that birds lack the gene PRDM9, which determines the location of recombination hotspots, and thus their recombination landscapes are more conserved than in other taxa, potentially allowing for clearer and longer term signals of gBGC (Singhal *et al.*, 2015).

Indeed, a number of studies have shown convincing evidence for the impact of gBGC in avian genomes. Recombination rates correlate with GC content of 3rd codon positions (GC3 content) in chicken (Rousselle *et al.*, 2019), GC content of fourfold degenerate sites (GC4 content) in flycatcher (Bolívar *et al.*, 2016) and with GC$*$ in zebra finch and long tailed finch (Singhal *et al.*, 2015) consistent with the action of gBGC. Additionally, Weber *et al.* (2014) recently made use of the large number of sequenced avian reference genomes to look at gBGC across the avian phylogeny, showing GC3 content correlates with a number of proxies for N_e, including body mass, longevity and female age at maturity, largely a result of elevated GC content in smaller species, consistent with stronger gBGC. Estimates of ancestral population sizes across this phylogeny also correlate with GC$*$ (Weber *et al.*, 2014).

Overall, GC content is increasing in birds (Webster, 2006), towards a higher equilibrium level, likely as a result of pervasive gBGC (Bolívar *et al.*, 2016; Nabholz *et al.*, 2011).

However, studies are largely focused on GC3 and GC4 and almost entirely confined to coding regions, and so may not be representative of the genome at large. Furthermore, work using ancestral repeat elements as a neutral reference suggests that there is a selective constraint of up to 40 percent on fourfold degenerate sites in birds (Kunstner *et al.*, 2011). This could be problematic for studies such as those discussed that rely on fourfold degenerate sites and third codon positions (which include all fourfold degenerate sites). Additionally, all of these approaches are based on substitution data, leaving the short-term dynamics unknown.

1.4 Quantifying Selection and the Strength of gBGC

The site frequency spectrum (SFS) is the population genetic workhorse for the analysis of polymorphism datasets. It summaries the allele frequencies of polymorphic variants in the population, providing both information on the mutation rate and selective forces operating. In its simplest form the SFS can be constructed from the frequencies of the least common allele of a variant (minor allele frequency). This results in a 'folded' SFS where the lower end of the distribution contains the variants segregating at high and low frequency in the population, and the upper end represents those at intermediate frequency. The folded SFS contains useful information for inferences about selection. For example, the classic measure of Tajima's D (Tajima, 1989) measures skews in the folded SFS, yielding negative values when it is enriched in low frequency and/or high frequency variants, which can be indicative of purifying or positive selection, whereas positive values reflect an enrichment of intermediate frequencies such as from balancing selection, and values close to zero reflect a neutral SFS. The folded SFS can also be used in more sophisticated quantification of the strength of selection through estimation of the distribution of fitness effects (DFE). The DFE is a means of categorising the selective pressures acting on new mutations, providing an estimate of the proportion of variants that are positively selected, neutral or deleterious within a population (see Eyre-Walker and Keightley, 2007, for review).

The folded SFS however is not always sufficient. The characterisation of the directional forces of selection on INDELs and gBGC on SNPs shares one methodological hurdle, namely the need to determine the ancestral states of the variants. Variant calling returns insertions and deletions relative to the reference genome, but without knowing if the longer variant or the shorter variant is ancestral, insertion events and deletion events cannot be teased apart. Similarly with gBGC, we are interested in estimating the strength of 'selection' for SW and WS polymorphisms separately, requiring knowledge of whether the S allele or the W allele is ancestral. In terms of the SFS this requires we obtain the derived allele frequency spectrum or 'unfolded' SFS. Inferring the ancestral state is relatively trivial conceptually, especially when using parsimony. Through the use of alignments between the study species and a number of out groups the most common state can be identified at a given position and taken to be ancestral. However, such approaches require good quality genome assemblies of closely related species and are prone to error. Errors in inferring the ancestral state, or 'polarisation errors' are particularly an issue for methods that use the unfolded site frequency spectrum to estimate population genetic parameters. This is due to the fact that a miss-orientated low frequency variant appears in the SFS as a high frequency variant, as well as the fact that for INDELs it also may switch mutation type from an insertion to a deletion, or vice versa which can confound estimates of selection (Hernandez *et al.*, 2007) (see figure 2.1 for a more detailed explanation). As such, until recently (e.g. Glémin *et al.*, 2015) these topics have largely been avoided, from the perspective of quantifying selection.

1.5 Thesis Chapters

In this thesis I take advantage of the public availability of high quality, whole genome resequencing datasets to address two understudied aspects of genome evolution, namely the quantification of selective pressures on small insertions and deletions and the impact of gBGC in the non-coding genome.

Chapter 2 presents a novel method for inferring the distribution of fitness effects from polymorphism data derived by Kai Zeng. The method is broadly applicable to combinations of SNP and INDEL data from a variety of genomic scales. The main advancement on previous methods is the possibility to derive the DFE for insertions and deletions separately whilst controlling for ancestral state misidentification. In this chapter I assess the model's performance with simulated datasets, before applying it to an INDEL dataset from publicly available resequenced *Drosophila melanogaster* genomes (Pool *et al.*, 2012). I demonstrate the model performs well, providing accurate results across a broad range of simulated parameters. My analysis of the *D. melanogaster* dataset reveals a bimodal DFE for INDELs in the coding regions of this species. Additionally I calculate the proportion of INDEL fixations that have been driven by positive selection (α) at $\sim$ 70 to 80%, a similar proportion to previously reported estimates for non-synonymous SNPs (Andolfatto *et al.*, 2011; Schneider *et al.*, 2011).

In Chapter 3 I investigate how natural selection has shaped INDEL variation in the genome of a wild passerine bird, the great tit (*Parus major*). I apply the model from Chapter 2 to an INDEL dataset derived from published resequencing data for 10 European great tits (Corcoran *et al.*, 2017). Analysis of coding sequence INDELs yields a bimodal DFE characterised by strong purifying selection, resulting in only $\sim$ 4% of INDEL events segregating in the population. Additionally, I estimate α at 71% and 86%, for insertions and deletions respectively. These results are in line with those reported in Chapter 2. Here I also extend my analysis to non-coding regions were I show that INDELs are still exposed to purifying selection but at a greatly reduced level, with $\sim$ 80% of insertions and $\sim$ 52% of deletions effectively neutral. I also show that in proximity to exons and in areas of low recombination INDEL diversity is reduced through the action of linked selection, and present some evidence for the mutagenic effect of recombination increasing INDEL mutation rates.

In Chapter 4, I shift my focus to the evolution of base composition in the great tit and an additional passerine, the zebra finch, focussing on the role of gBGC and INDELs. I generate a dataset of 1 megabase orthologous windows of non-coding data from the great tit genomes used in Chapter 2 and 10 zebra finch genomes from Singhal *et al.* (2015). I estimate the population scaled strength of gBGC (B) using the model from

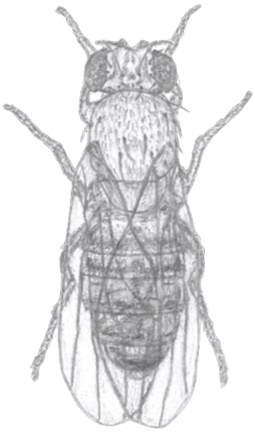
Glémin *et al.* (2015) and assess gBGC's contribution to the base composition of the two species' genomes. I also evaluate the impact of small insertions and deletions on the genomes' base composition. The analysis demonstrates remarkable conservation in the underlying strength of conversion bias, with increased B estimates in the zebra finch of the same magnitude of the species' larger N_e. Overall, the analysis shows non-coding B values are weak (< 1), and INDELs have not been GC conservative in their impact on the lineages leading to the two species.

Chapter 2

New methods for inferring the distribution of fitness effects for INDELs and SNPs

Authors: **Henry J. Barton** and **Kai Zeng**

This chapter has been published in the *Molecular Biology and Evolution* under the same title, see Barton and Zeng (2018). It is shown here in its published form with additional formatting changes.



2.1 Abstract

Small insertions and deletions (INDELs; ≤50bp) are the most common type of variability after SNPs. However, compared to SNPs, we know little about the distribution of fitness effects (DFE) of new INDEL mutations and how prevalent adaptive INDEL substitutions are. Studying INDELs has been difficult partly because identifying ancestral states at these sites is error-prone and misidentification can lead to severely biased estimates of the strength of selection. To solve these problems, we develop new maximum likelihood methods, which use polymorphism data to simultaneously estimate the DFE, the mutation rate, and the misidentification rate. These methods are applicable to both INDELs and SNPs. Simulations show that they can provide highly accurate results. We applied the methods to an INDEL polymorphism dataset in *Drosophila melanogaster*. We found that the DFE for polymorphic INDELs in protein-coding regions is bimodal, with the variants being either nearly neutral or strongly deleterious. Based on the DFE, we estimated that 71.5% – 83.7% of the INDEL substitutions that took place along the *D. melanogaster* lineage were fixed by positive selection, which is comparable to the prevalence of adaptive substitutions at non-synonymous sites. The new methods have been implemented in the software package `anavar`.

2.2 Introduction

New mutations can have a range of effects on an organism's fitness, ranging from being strongly harmful, through being only slightly deleterious, to being neutral, and finally on to being either mildly or highly beneficial. The relative frequencies of mutations with different selective effects is known as the distribution of fitness effects (DFE). The DFE is an important parameter as it is required for addressing many fundamental questions (Eyre-Walker and Keightley, 2007). Examples include understanding determinants of the efficacy of natural selection (Corcoran *et al.*, 2017; Galtier, 2016), the genetic basis of polygenic traits (Zuk *et al.*, 2014), and the evolutionary advantage of sex and recombination (Hartfield and Keightley, 2012).

Taking advantage of the massive increase in data availability, many methods have been proposed for estimating the DFE using polymorphism data (Eyre-Walker and Keightley, 2009; Eyre-Walker *et al.*, 2006; Keightley *et al.*, 2009; Kim *et al.*, 2017; Kousathanas and Keightley, 2013; Tataru *et al.*, 2017). Their development in turn allows more reliable inferences about other important quantities such as α, the proportion of adaptive substitutions (Eyre-Walker and Keightley, 2009). However, all these methods are concerned with estimating the DFE for single nucleotide polymorphisms (SNPs). Consequently, much less is known about the DFE and α for other types of genetic variation such as small insertions and deletions (INDELs; $\leq$ 50bp), despite the fact that INDELs are the second most common type of variants (e.g., Montgomery *et al.*, 2013), and hence represent an important source of raw materials for selection to act on.

A major difficulty in studying INDELs lies with ancestral state identification. This requires multi-species genome alignments. However, INDELs occur disproportionately in repetitive genomic regions (Ananda *et al.*, 2013; Montgomery *et al.*, 2013), where alignment algorithms perform poorly (Earl *et al.*, 2014). Furthermore, there is evidence that homoplasy is a significant issue outside repetitive regions, probably due to the existence of cryptic INDEL mutation hotspots (Kvikstad and Duret, 2014). Thus ancestral state identification can be expected to be particularly error prone for INDELs. It is well established that misidenfication of ancestral states can lead to severely biased estimates of the strength of selection using the site-frequency spectrum (SFS) (Hernandez *et al.*, 2007). For SNPs, this difficulty can be avoided by using the folded SFS (e.g., Eyre-Walker *et al.*, 2006; Keightley and Eyre-Walker, 2007). However, to determine whether a length variant is an insertion or a deletion, we have to know what the ancestral state is, meaning that the issue of polarisation error is inherent for INDELs. As a result, applying existing methods for estimating the DFE to INDEL data may be liable to biases.

Another challenge is that the SFSs for insertions and deletions may be affected by polarisation errors to different extents. This is because when the ancestral state of an insertion segregating at low frequency is misidentified, it will be incorrectly inferred as a deletion segregating at high frequency (and vice versa). There is direct experimental

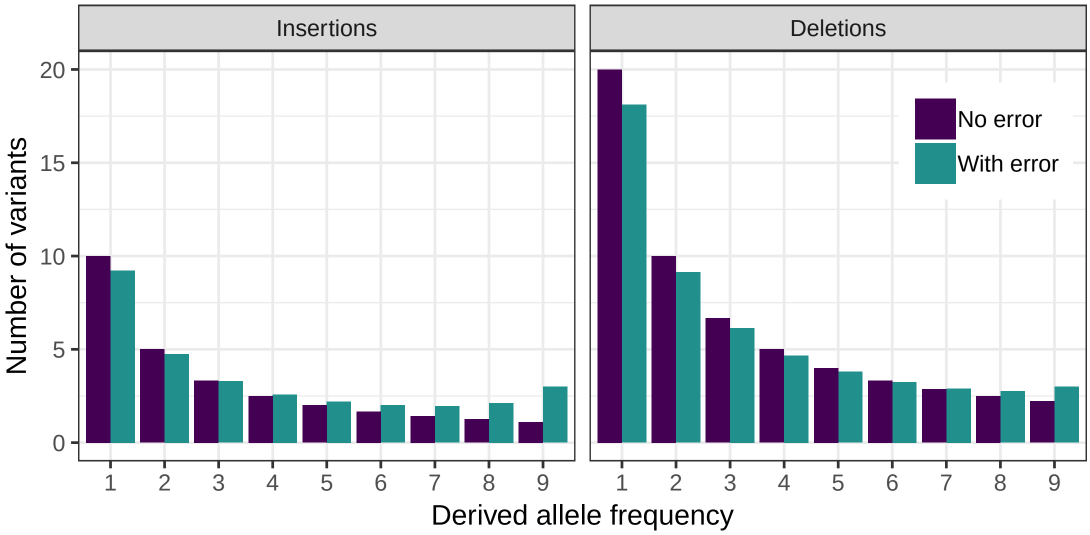


FIGURE 2.1: The SFSs for insertions and deletions may be affected to different extents by polarisation errors. We assume that the population size is constant, that INDELs are neutral, and that the sample size is 10. In the genomic region under consideration, the total scaled mutation rate towards insertions, $4N_eum$, is 10, where N_e is the effective population size u is the insertion mutation rate per site per generation, and m is that size of the focal region. The total scaled mutation rate towards deletions is 20. The expected SFSs were generated using standard neutral theory. The SFSs with polarisation errors were generated by assuming that the ancestral state of an INDEL was wrongly identified with probability 0.1.

evidence that the deletion mutation rate is higher than the insertion mutation rate (Besenbacher *et al.*, 2015; Keightley *et al.*, 2009; Schrider *et al.*, 2013; Yang *et al.*, 2015). This mutational bias means that there are more deletions segregating in the population than insertions. The larger number of deletions may lead to the SFS for insertions being disproportionally affected by polarisation errors (Figure 2.1). This asymmetry can cause the insertion SFS to have a more pronounced, but artificial, uptick at the high-frequency end, which can be misinterpreted as stronger positive selection on insertions over deletions. As pointed out by Kvikstad and Duret (2014), this methodological issue can, at least in principle, compromises the results of previous studies, which suggest that insertions are more likely to be under positive selection than deletions to prevent the genome size from unconstrained contraction caused by the mutational bias towards deletions (Parsch, 2003). Similarly, it will make it difficult to test the possibility that insertions have a higher fixation probability because they are favoured by insertion-biased gene conversion (Leushkin and Bazykin, 2013).

Towards resolving the confounding efforts ancestral state misidentification have on the study of INDELs, we propose new maximum likelihood methods for inferring the DFE using polymorphism data. These methods are based on recent studies on SNPs which show that polymorphism data contains enough information for simultaneous estimation of the mutation rate, the DFE, and the polarisation error rate (Glémin *et al.*, 2015; Tataru *et al.*, 2017). Our methods are more general than the existing methods in the following aspects. First, they can handle both INDELs and SNPs. Second, insertions and deletions can have different polarisation error rates, mutation rates, and DFEs. Third, for both INDELs and SNPs, the new methods allow the mutation and polarisation error rates to vary across the genome. Incorporating these heterogeneities may be particularly important for INDELs (Kvikstad and Duret, 2014). We carried out extensive simulations to examine the performance of the new methods. As an example, we applied the methods to an INDEL polymorphism dataset in *Drosophila melanogaster* we obtained by re-analysing the raw short-read data published by the *Drosophila* Population Genomics Project (Pool *et al.*, 2012). Through model comparisons, we tried to find the DFE that best described the observed pattern of INDEL polymorphism within protein-coding regions of the genome. Finally, using the best-fitting DFE, we estimated the proportion of INDEL substitutions fixed by positive selection (α).

2.3 New Approach

For ease of presentation, we will start with a description of the SNP models. The INDEL models will be presented later as an extension.

2.3.1 The SNP models

Consider a diploid population with effective size N_e. The size of the genomic region of interest is m base pairs, and the sample size is n.

2.3.1.1 The discrete model:

Assume that there are C different classes of sites in the focal region. These sites can be different with respect to their mutation rates, the fitness effects of new mutations, and polarisation error rates. This discrete model has several advantages. First, it does not assume that the DFE follows a specific probability distribution, and is therefore able to accommodate complex scenarios such as a multi-modal DFE (Kousathanas and Keightley, 2013). Second, by allowing the mutation and polarisation error rates to vary freely between site classes, the method can include situations whereby these two variables co-vary (e.g., hypermutable regions may have a higher polarisation error rate).

We assume that the mutation process can be approximated by the infinite-sites model. Let the total scaled mutation rate for sites of class c be $m\theta_c$, where $c \in \{1, 2, ..., C\}$ and $\theta_c = 4N_e u_c$. To understand u_c, consider an alternative formulation whereby the mutation rate for the c^{th} class of sites is v_c per site per generation, and sites of class c account for a fraction p_c of all sites in the focal region (i.e., $\sum_c p_c = 1$). We have $m\theta_c = mp_c 4N_e v_c$, which leads to $u_c = p_c v_c$. By using θ_c, we can perform searches for maximum likelihood estimates (MLEs) of the parameters without having to deal with the constraint $\sum_c p_c = 1$. Define

$$\theta = \sum_{c=1}^{C} \theta_c = 4N_e \sum_{c=1}^{C} p_c v_c. \tag{2.1}$$

Thus, θ is the average scaled mutation rate per site, and the total scaled mutation rate is $m\theta$. If the per-site mutation rate is uniform across the focal region (i.e., $v_i = v_j$ for $i \neq j$ and $1 \leq i, j \leq C$), then $\theta_c/\theta = p_c$.

To model selection, we assume that, for mutations arising at sites of class c, the fitnesses of the wild-type, heterozygote, and mutant homozygote genotypes are 1, 1 + s_c, and 1 + $2s_c$, respectively. The corresponding scaled selection coefficient γ_c is defined as $4N_e s_c$. Positive and negative γ_c values signify beneficial and deleterious mutations, respectively.

The site-frequency spectrum (SFS) for the c^{th} site class, which is defined as the expected number of polymorphic sites of size i (i.e., sites where the derived allele is represented i

times; $1 \leq i < n$), is given by

$$\Psi_{c,i} = m\theta_c \tau_i(\gamma_c) \tag{2.2}$$

where

$$\tau_i(\gamma) = \int_0^1 \binom{n}{i} x^i (1-x)^{n-i} \frac{1 - e^{-\gamma(1-x)}}{x(1-x)(1-e^{-\gamma})} dx. \tag{2.3}$$

Polarisation errors distort the SFS. Specifically, when the ancestral state of a polymorphic site of size i is mis-identified, it will be regarded as a polymorphic site of size $n-i$. To model polarisation errors, we let ϵ_c be the probability that the ancestral state of a polymorphic site of class c is incorrectly identified (Glémin *et al.*, 2015). The final SFS for sites of class c is then

$$\psi_{c,i} = (1-\epsilon_c)\Psi_{c,i} + \epsilon_c \Psi_{c,n-i}. \tag{2.4}$$

In what follows, we refer to the SFS with and without the correction of polarisation errors as the corrected and uncorrected SFS, respectively. The corrected SFS for the focal region is simply the sum of all the contributions from the sites in different classes

$$\psi_i = \sum_{c=1}^{C} \psi_{c,i}. \tag{2.5}$$

Existing models either do not model polarisation error (Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2007; Kim *et al.*, 2017) or assume that the error rate is constant across the focal region (Glémin *et al.*, 2015; Tataru *et al.*, 2017). The model described above is therefore more general. Allowing variation in the polarisation error rate can be important. For instance, sites under stronger selective constraints tend to evolve slower, and are less likely to be polarised incorrectly due to homoplasy. It should, however, be noted that, when $\gamma_c \equiv \gamma$ for $\forall c \in \{1, 2, ..., C\}$, not all the parameters are identifiable. To see this, we rewrite (2.5) as

$$\psi_i = m \sum_{c=1}^{C} (1-\epsilon_c)\theta_c \tau_i(\gamma) + m \sum_{c=1}^{C} \epsilon_c \theta_c \tau_{n-i}(\gamma). \tag{2.6}$$

Appealing to (2.1) and defining ϵ^* such that

$$\epsilon^*\theta = \sum_{c=1}^{C} \epsilon_c \theta_c \tag{2.7}$$

we can rewrite (2.6) as

$$\psi_i = (1 - \epsilon^*) m\theta\tau_i(\gamma) + \epsilon^* m\theta\tau_{n-i}(\gamma). \tag{2.8}$$

Thus, when there is no difference in fitness effects between mutations arising at sites of different classes, we cannot detect variation in the scaled mutation rate and polarisation error rate because the model reduces to one that depends on θ, γ and ϵ^*. This result has important implications for data analysis by pointing out that a model with a small number of site classes may provide an adequate description of the data even when the underlying biological process features complex variation in the mutation rate across the genome.

2.3.1.2 The continuous model:

Instead of assuming that the focal region is composed of several classes of sites, we can assume that the fitness effects of new mutations follows a continuous distribution characterised by parameters Ω. Let θ be the scaled mutation rate per site, and ϵ be the polarisation error rate. The uncorrected SFS becomes

$$\Psi_i = m\theta \int \tau_i(\gamma) f(\gamma|\Omega) d\gamma \tag{2.9}$$

where $f(\gamma|\Omega)$ is the probability density function. The corrected SFS is analogous to (2.4) with c in the subscripts omitted.

Although the modelling framework allows the DFE to follow arbitrary probability distribution (including those mixture distributions considered by Galtier (2016)), here we only consider the reflected Γ distribution, i.e., $-\gamma \sim \Gamma(a, b)$, where $\gamma \leq 0$ and a and b are the shape and scale parameters, respectively.

2.3.1.3 Parameter estimation:

Let $X = (x_1, x_2, ..., x_{n-1})$ represent the observed SFS, where x_i is the number of polymorphic sites of size i in the sample. Let Θ denote all the parameters in the model (i.e., θ_c, γ_c, and ϵ_c for $c \in \{1, 2, ..., C\}$ for the discrete model and θ, Ω, and ϵ for the continuous model). To obtain MLEs of Θ, we use the Poisson random field model (Bustamante *et al.*, 2001; Sawyer and Hartl, 1992). Omitting constants that have no effects on the shape of the likelihood surface, the log likelihood function is defined as

$$L(\Theta|X) = \sum_{i=1}^{n-1} \big(-\psi_i + x_i \ln(\psi_i) \big). \tag{2.10}$$

2.3.1.4 Controlling for demography:

We have so far assumed that the population is panmictic and of constant size N_e. To control for demography, we employ the method of Eyre-Walker *et al.* (2006). Take the continuous model as an example. First, we define augmented SFSs as

$$\begin{cases} \Psi_i^* = r_i \Psi_i & \text{(2.11a)} \\ \psi_i^* = (1-\epsilon)\Psi_i^* + \epsilon \Psi_{n-i}^* & \text{(2.11b)} \end{cases}$$

Next, a set of neutral variants is added to the model, which introduces two additional parameters $\theta^{(0)}$ and $\epsilon^{(0)}$, which are the scaled mutation rate per site and the polarisation error rate, respectively, for the neutral sites. Let $\Theta^{(0)}$ denote these new parameters and $X^{(0)}$ denote the neutral SFS. The log likelihood of the observed data can be calculated as

$$L(\Theta, \Theta^{(0)}, R|X, X^{(0)}) = L(\Theta, R|X) + L(\Theta^{(0)}, R|X^{(0)}) \tag{2.12}$$

where $R = (r_2, r_3, ..., r_{n-1})$ and the two log likelihood functions on the right-hand side are calculated in the same way as (2.10) with ψ_i replaced by ψ_i^*.

The above method for controlling for demography has been used extensively (Eyre-Walker *et al.*, 2006; Galtier, 2016; Glémin *et al.*, 2015; Jackson *et al.*, 2017; Muyle *et al.*, 2011; Tataru *et al.*, 2017). These previous efforts have gathered clear theoretical and empirical evidence that the method is robust against a wide range of demographic processes, as well as the effects caused by selection at linked sites (e.g., background selection and/or selective sweeps). For instance, in a recent analysis of selection on codon usage bias in *Drosophila*, Jackson *et al.* (2017) showed that the estimates of γ produced by an estimation method that corrects for demography using the r parameters as set out above closely matched those produced by another estimation method that considers an explicit one-step change in population size (see Figure 4A in Jackson *et al.* (2017)).

It should be noted that (2.12) accommodates the possibility that the focal region and the neutral region have different mutation rates. This is more general than several previous models (Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2007; Kim *et al.*, 2017; Tataru *et al.*, 2017). However, it may be challenging to distinguish this model from one in which the two regions have the same mutation rate, but a proportion of new mutations in the focal region are so strongly deleterious that they make negligible contributions to the observed SFS.

2.3.2 The INDEL models

2.3.2.1 The discrete model:

First consider insertions. Assume that there are C^{ins} different classes of sites. The total scaled mutation rate towards insertions for sites of class c is $m\theta_c^{ins}$, and the fitness effect and polarisation error rate are γ_c^{ins} and ϵ_c^{ins}, respectively ($1 \leq c \leq C^{ins}$). The uncorrected SFS for insertions of class c can be calculated using (2.2), and is denoted by $\Psi_{c,i}^{ins}$. For deletions, we can similarly assume that there are C^{del} different classes of sites. The associated parameters are θ_d^{del}, γ_d^{del}, and ϵ_d^{del}, and the uncorrected SFS is denoted by $\Psi_{d,i}^{del}$ ($1 \leq d \leq C^{del}$).

When the ancestral state of a derived insertion of size i is misidentified, it will be wrongly identified as a deletion of size $n - i$, and vice versa for deletions (note that size in this context refers to the frequency of the derived allele, not the number of base pairs inserted or deleted). Thus, the corrected SFSs for insertions and deletions are

$$\begin{cases} \psi_i^{ins} = \sum_{c=1}^{C^{ins}} (1-\epsilon_c^{ins})\Psi_{c,i}^{ins} + \sum_{d=1}^{C^{del}} \epsilon_d^{del}\Psi_{d,n-i}^{del} & \text{(2.13a)} \\ \psi_i^{del} = \sum_{d=1}^{C^{del}} (1-\epsilon_d^{del})\Psi_{d,i}^{del} + \sum_{c=1}^{C^{ins}} \epsilon_c^{ins}\Psi_{c,n-i}^{ins} & \text{(2.13b)} \end{cases}$$

2.3.2.2 The continuous model:

For insertions, define the per-site scaled mutation rate and the polarisation error rate as θ^{ins} and ϵ^{ins}, respectively. The DFE for insertions is determined by parameters Ω^{ins}. For deletions, we similarly define the following parameters: θ^{del}, Ω^{del} and ϵ^{del}. Finally, the corrected SFSs are

$$\begin{cases} \psi_i^{ins} = (1-\epsilon^{ins})\Psi_i^{ins} + \epsilon^{del}\Psi_{n-i}^{del} & \text{(2.14a)} \\ \psi_i^{del} = (1-\epsilon^{del})\Psi_i^{del} + \epsilon^{ins}\Psi_{n-i}^{ins} & \text{(2.14b)} \end{cases}$$

where Ψ_i^{ins} and Ψ_i^{del} are the uncorrected SFSs for insertions and deletions, respectively, and are calculated in the same way as (2.9). As in the SNP case, we only consider cases where the DFE follows a reflected Γ distribution. The shape and scale parameters for insertions and deletions are denoted by a^{ins}, b^{ins}, a^{del}, and b^{del}, respectively.

2.3.2.3 Parameter estimation:

Let $X^{ins} = (x_1^{ins}, x_2^{ins}, ..., x_{n-1}^{ins})$ and $X^{del} = (x_1^{del}, x_2^{del}, ..., x_{n-1}^{del})$ be the observed SFSs for insertions and deletions, respectively. The log likelihood of the data is calculated as

$$L(\Theta|X^{ins}, X^{del}) = \sum_{z\in\{ins,\ del\}} \sum_{i=1}^{n-1} \big(-\psi_i^z + x_i^z \ln(\psi_i^z)\big). \tag{2.15}$$

2.3.2.4 Controlling for demography:

Take the continuous model as an example. The augmented SFSs are

$$\begin{cases} \Psi_i^{ins,*} = r_i \Psi_i^{ins} & (2.16a) \\ \Psi_i^{del,*} = r_i \Psi_i^{del} & (2.16b) \\ \psi_i^{ins,*} = (1-\epsilon^{ins})\Psi_i^{ins,*} + \epsilon^{del}\Psi_{n-i}^{del,*} & (2.16c) \\ \psi_i^{del,*} = (1-\epsilon^{del})\Psi_i^{del,*} + \epsilon^{ins}\Psi_{n-i}^{ins,*} & (2.16d) \end{cases}$$

As for the neutral reference, we can in principle use any combinations of SNPs, insertions, and deletions collected from putatively neutrally evolving regions. Assume that we have access to both neutral insertions and neutral deletions, and the observed SFSs are denoted by $X^{ins,(0)}$ and $X^{del,(0)}$, respectively. The additional parameters needed to model the neutral variants include $\theta^{ins,(0)}$, $\epsilon^{ins,(0)}$, $\theta^{del,(0)}$, and $\epsilon^{del,(0)}$, which are denoted collectively by $\Theta^{(0)}$. The log likelihood is

$$\begin{aligned} &L(\Theta, \Theta^{(0)}, R|X^{ins}, X^{del}, X^{ins,(0)}, X^{del,(0)}) \\ &= L(\Theta, R|X^{ins}, X^{del}) + L(\Theta^{(0)}, R|X^{ins,(0)}, X^{del,(0)}) \qquad (2.17) \end{aligned}$$

where the two terms on the right are calculated using (2.15) with ψ_i^z replaced by $\psi_i^{z,*}$ ($z \in \{ins,\ del\}$).

2.4 Results and Discussion

2.4.1 Simulation results

We evaluate the statistical properties of the new models using computer simulations. Unless stated otherwise, the sample size (n) is 50 and the results are based on 100 replicates. In all cases, we assume the population size is constant and only analyse data from the selected region (see Materials and Methods for justification). For the SNP models, we only present results for the discrete SNP model with $C > 1$ site classes,

Table 2.1: Maximum likelihood estimates (MLEs) of the parameters of discrete SNP models with $C = 2$ classes of sites. Simulated data were generated using the parameter values shown in the "True value" rows, with two different region sizes, m. For each parameter combination, 100 samples of size 50 were simulated and analysed to obtain MLEs.

	m	C	θ	γ	ϵ
True value	–	1	0.005	-5	0.05
		2	0.01	-20	0.01
Mean (SD) of MLEs	10^6	1	0.0050 (0.0007)	-5.0 (0.4)	0.051 (0.006)
		2	0.010 (0.001)	-20.2 (1.9)	0.009 (0.006)
	10^5	1	0.0044 (0.0017)	-4.4 (1.5)	0.042 (0.022)
		2	0.011 (0.001)	-20.0 (5.7)	0.016 (0.014)

because both the $C = 1$ case and the continuous model have been analysed before (Glémin *et al.*, 2015; Tataru *et al.*, 2017).

2.4.1.1 Properties of the discrete SNP model:

First consider a model with $C = 2$ site classes. As can be seen from Table 2.1, there is information in the SFS for simultaneously estimating all the parameters to a high degree of accuracy. Before discussing more simulation results, it should be pointed out that, when $C > 1$, the order of the site classes is arbitrary. That is, the model considered in Table 2.1 is equivalent to one with parameters $\theta_1 = 0.01$, $\gamma_1 = -20$, $\epsilon_1 = 0.01$, $\theta_2 = 0.005$, $\gamma_2 = -5$, and $\epsilon_2 = 0.05$. For both cases shown in Table 2.1, all the MLEs can be sorted such that $\hat{\theta}_1 < \hat{\theta}_2$ and $\hat{\gamma}_1 > \hat{\gamma}_2$. In other words, the MLEs can be assigned unambiguously to site classes according to the order given in the "True value" row. However, if we were to reduce the amount of data, parameter estimates will become more uncertain, and cases such as those with $\hat{\theta}_1 < \hat{\theta}_2$ and $\hat{\gamma}_1 < \hat{\gamma}_2$ will occur, which makes assigning the MLEs to site classes impossible. Thus, presenting mean and standard deviation of the MLEs may give misleading information about the performance of the model.

In light of the above discussion, we investigate the statistical properties of the model using two alternative methods. First, we compare the full model to the following reduced models using the χ^2 test: "Equal ϵ" (all site share the same polarisation error rate), "$\epsilon = 0$" (no polarisation error), and "$C - 1$" (a model with $C - 1$ site classes, where C is the true number of site classes). Second, we assess how well these various models

Table 2.2: Statistical properties of the discrete SNP model. The parameters used in Case 3 were $\theta_1 = 0.002$, $\gamma_1 = 0$, $\epsilon_1 = 0.05$, $\theta_2 = 0.006$, $\gamma_2 = -5$, $\epsilon_2 = 0.02$, $\theta_3 = 0.002$, $\gamma_3 = -30$, $\epsilon_3 = 0.01$, and $n = 100$. A large sample size was used for Cases 3 and 4 due to the inclusion of strongly deleterious mutations (i.e., $\gamma_3 = -30$). Values under "Percent significant" show how often the full model fitted the data better than the three reduced models (see the main text for more details). The $\bar{\mu}$ (see (2.18) in Materials and Methods) obtained under the $\epsilon = 0$ model are large because ignoring polarisation error results in the inference of a site class with a strongly positive γ.

Case	Parameters	m	Percent significant			$\bar{\mu}$				
			Equal ϵ	$\epsilon = 0$	$C-1$	True	Full	Equal ϵ	$\epsilon = 0$	$C-1$
1	Same as Table 2.1	10^6	93	100	100	0.0113	0.0114	0.0171	> 1	0.0022
2	Same as Table 2.1	10^5	15	92	100	0.0113	0.0158	0.0204	> 1	0.0022
3	See legend above	10^7	3	100	100	0.2204	0.2267	0.2613	> 1	0.1755
4	Same as Case 3	2×10^6	0	33	55	0.2204	0.2271	0.2580	> 1	0.1768

predict the average fixation probability $\bar{\mu}$ (see (2.18) in Materials and Methods), which is essential for estimating the prevalence of adaptive substitutions (i.e., α and ω_a).

Considering the two pairs of cases in Table 2.2, and focusing on the data presented under "Percent significant", we make the following observations. First, as the amount of data reduces, the ability of the model to infer separate ϵ for different site classes drops more rapidly than its ability to detect the existence of either polarisation error or more than one site class. This suggests that estimating heterogeneity in ϵ may be challenging. Considering all four cases, it appears that the tests for detecting the presence of polarisation error (i.e., the full model versus "$\epsilon = 0$") and for detecting the existence of more site classes (i.e., the full model versus "$C - 1$") are more powerful, especially the latter. It should be noted that the likelihood surface appears to be rather flat when $C = 3$ such that different parameter combinations may produce very similar log likelihoods. This is particularly evident when the amount of data is limited (Case 3 versus Case 4), leading to a reduction in power of the tests. A similar observation was made by Keightley and Eyre-Walker (2010), who also showed that it can be partly alleviated by increasing the sample size. Nonetheless there may well be a limit as to how many site classes can be included. This identifiability problem is analogous to that discussed extensively in the context of using SNP-based methods for estimating past demographic changes (e.g., Myers *et al.*, 2008).

Interestingly, the reduced model "Equal ϵ" makes worse predictions of $\bar{\mu}$ than the full model in all cases presented in Table 2.2, even when the full model does not normally

Table 2.3: MLEs of the parameters of several INDEL models

Model	m	Parameters								
Discrete	2×10^6	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}		
		True	0.0005	-5	0.02	0.001	-15	0.02		
		Mean MLE	0.00050	-5.0	0.021	0.0010	-15.0	0.020		
Continuous	2×10^7	Name	θ^{ins}	a^{ins}	b^{ins}	ϵ^{ins}	θ^{del}	a^{del}	b^{del}	ϵ^{del}
		True	0.0005	0.5	10	0.08	0.001	0.25	50	0.04
		Mean MLE	0.00050	0.51	10.4	0.080	0.0010	0.251	51.2	0.040
Continuous	2×10^6	Name	θ^{ins}	a^{ins}	b^{ins}	ϵ^{ins}	θ^{del}	a^{del}	b^{del}	ϵ^{del}
		True	0.0005	0.5	10	0.08	0.001	0.25	50	0.04
		Mean MLE	0.00054	0.51	144.7	0.082	0.0010	0.253	93.2	0.041

provide a better fit to the data (Cases 2 and 4). The same applies to the other two reduced models. Thus, despite the statistical difficulties discussed above, fitting the full model to the data may be important for obtaining accurate estimates of α and ω_a.

2.4.1.2 Properties of the INDEL models:

Table 2.3 contains simulation results based on a discrete model (with $C^{ins} = C^{del} = 1$) and two continuous models (differing from each other in terms of the size of the focal region m). The mutation rates are about 10 times lower than those used in the SNP cases (Tables 2.1 and 2.2), and polarisation error rates are about 2 times higher. These choices are to reflect the fact that INDELs are generally less prevalent than SNPs, and are potentially more difficult to polarise. As can be seen, with a reasonable amount of data, all the parameters can be reliably estimated. Comparing the two continuous models, we notice that, with limited data, the scale parameter b of the Γ distribution may be overestimated, but estimates of the shape parameter a and the polarisation error rate remain unbiased.

The true values of $\bar{\mu}^{ins}$ and $\bar{\mu}^{del}$ for the discrete model are 0.0339 and 4.59×10^{-6}, respectively. The mean (SD) of the estimates is 0.0345 (0.0055) for $\bar{\mu}^{ins}$, and 5.27×10^{-6} (2.91×10^{-6}) for $\bar{\mu}^{del}$. Thus, the true values are well within the observed ranges of variability. The true values of $\bar{\mu}^{ins}$ and $\bar{\mu}^{del}$ for the two continuous cases are 0.384 and 0.429, respectively. The mean (SD) of the estimates for the case with more data is 0.382 (0.012) for $\bar{\mu}^{ins}$ and 0.429 (0.008) for $\bar{\mu}^{del}$. Encouragingly, for the continuous case with less data, despite the tendency to overestimate the scale parameter, estimates of the

Table 2.4: Summary statistics for the INDEL and SNP data

Data	Type	Diversity (π)	Tajima's D
INDELs	CDS	5.20×10^{-5}	-1.208
	Frameshift	2.06×10^{-5}	-1.253
	Non-frameshift	3.14×10^{-5}	-1.177
	Intron	0.0016	-0.729
	Intergenic	0.0017	-0.704
	Non-coding	0.0017	-0.718
SNPs	Nonsense	5.83×10^{-6}	-1.510
	0-fold degenerate sites	0.0016	-0.868
	4-fold degenerate sites	0.0165	-0.210

average fixation probabilities are still highly accurate: 0.388 (0.050) for $\bar{\mu}^{ins}$ and 0.418 (0.028) for $\bar{\mu}^{del}$, suggesting that the reliability of estimates of α and ω_a is unlikely to be compromised.

2.4.2 Application to *D. melanogaster* data

2.4.2.1 A summary of the data

Using the variant calling pipeline detailed in Materials and Methods, a total of 370,217 INDELs ($\leq$ 50bp) and 1,789,367 SNPs were identified from the 17 Rwandan individuals. Our analysis primarily focuses on INDELs because SNPs have been analysed extensively before (Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2007; Schneider *et al.*, 2011). Similar to previous reports (e.g., Ptak and Petrov, 2002), smaller INDELs are more prevalent than larger ones (supplementary Figure A.1). INDEL diversity is about 30 times lower in protein-coding (CDS) regions than in either intronic or intergenic regions (Table 2.4). Additionally, frameshift INDELs are rarer than non-frameshift ones (Table 2.4; supplementary Figure A.1). Interestingly, nonsense mutations are somewhat rarer than frameshift INDELs, an observation also made by Leushkin *et al.* (2013). These results indicate strong purifying selection against INDELs in protein-coding regions. INDEL diversity patterns appear to be similar between intronic and intergenic regions. They are combined and referred to as non-coding INDELs in what follows to increase statistical power.

Comparing between INDELs and SNPs, we notice that INDEL diversity in non-coding regions is about 10 times lower than π_4 (4-fold site diversity; Table 2.4), consistent with the fact that the INDEL mutation rate is lower than the point mutation rate (Haag-Liautard *et al.*, 2007; Schrider *et al.*, 2013). However, Tajima's D calculated on non-coding INDELs is more negative than that calculated on 4-fold sites (Table 2.4), probably reflecting the fact that many non-coding DNA in the *D. melanogaster* genome are under selection (Andolfatto, 2005). Furthermore, π_0 (0-fold site diversity; Table 2.4) is only about 10 times smaller than π_4. This level of reduction is much smaller than the 30-fold difference observed between CDS and non-coding INDELs. This suggests that, in protein-coding regions, INDEL mutations are under much stronger purifying selection than 0-fold mutations, which is consistent with the more negative Tajima's D value calculated on CDS INDELs (Table 2.4).

To further investigate the data, we calculated d_N, substitution rate at nonsynonymous sites, using PAML and the reference genomes of *D. simulans* and *D. yakuba* (see Materials and Methods). The genes were then divided into 20 equal-sized bins. For each bin, we calculated average π_0 and π_{INDEL}. Both statistics decrease as d_N decreases (Figure A.2), consistent with the expectation that mutations are on average more deleterious in more conserved genes (Jackson *et al.*, 2015). The results in this and the preceding paragraphs suggest that our INDEL dataset is of high quality.

2.4.2.2 Inferring the DFE and α using non-coding INDELs as the neutral reference

To infer the DFE for INDELs in CDS regions, we used non-coding INDELs as the neutral reference. Following previous efforts in estimating the DFE for SNPs (Eyre-Walker and Keightley, 2009; Galtier, 2016; Keightley and Eyre-Walker, 2007; Schneider *et al.*, 2011; Tataru *et al.*, 2017), we also assumed that the mutation rate towards insertions and deletions, respectively, were the same between the neutral and selected regions. The best-fitting DFE is one with $C = 2$ classes of selected sites (Table 2.5 and supplementary Table A.1). The MLEs of γ suggest that polymorphic INDELs are either nearly neutral or are so strongly deleterious that they contribute little to polymorphism. This seems to

TABLE 2.5: Results based on the best-fitting models for INDELs in the CDS regions of the *D. melanogaster* genome. The DFE for polymorphic INDELs in the CDS regions were inferred using either non-coding INDELs or 4-fold sites as the neutral reference. A series of different DFEs were fitted to the data, and the best-fitting models presented above were determined by using the Akaike information criterion (AIC) (see supplementary Tables S1 and S3). When non-coding INDELs were used as the neutral reference, α was estimated using INDEL divergence in noncoding regions. When 4-fold sites were used as the neutral reference, the mutation rate ratio between SNPs and INDELs, and that between deletions and insertions, were fixed at values obtained from a mutation accumulation experiment (Schrider *et al.*, 2013). α was estimated using a method based on divergence in the 8–30bp region of short introns < 66bp long (see the main text).

Model	Parameters for CDS INDELs						α
Noncoding INDELs	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	83.7%
Discrete $C = 2$	1.8×10^{-5}	1.98	0.023	5.3×10^{-5}	-1.69	0.016	
Uniform mutation rate	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	7.2×10^{-4}	-1566.4	3.6×10^{-5}	0.0011	-642.5	1.6×10^{-5}	
4-fold degenerate sites	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	71.5%
Discrete $C = 2$	1.6×10^{-5}	-1.31	0.0092	4.9×10^{-5}	-3.77	0.0082	
Fixed mutation ratios	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	1.9×10^{-4}	-284.1	1.2×10^{-4}	0.0010	-454.8	6.2×10^{-5}	

be consistent with the 30-fold difference in INDEL diversity level between CDS and noncoding regions, which is more substantial than the 10-fold difference between 0-fold and 4-fold sites (Table 2.4). Fitting the data to a discrete model with $C = 3$ classes of sites also reveals a bimodal DFE, suggesting that the conclusion is robust (supplementary Table A.1). With a larger sample containing hundreds or even thousands of alleles, and by fitting a DFE with more site classes, it should be possible to obtain further details of the relative frequencies and fitness effects of strongly selected variants, which tend not to segregate in our current sample of size 17. However, this additional information about the strongly selected end of the DFE is unlikely to affect our estimation of α (see below) because these variants make effectively no contribution to divergence.

To better understand the effects of length, we separated the INDELs in CDS regions into the following length categories: 1bp, 2bp, 3bp, frameshifting (≥4bp), and non-frameshifting (≥6bp). We analysed the data in each category separately. As above, noncoding INDELs with the same length were used as the neutral reference and the mutation rate was assumed to be constant across neutral and selected sites. Considering the dearth of variants, we only fitted a DFE with $C = 1$ class of selected sites. Viewing the γ in this model as the "average" selection coefficient, frameshift INDELs are consistently more deleterious than non-frameshift INDELs (supplementary Figure A.3). Consistent with a prevous study (Leushkin *et al.*, 2013), there is no obvious evidence that longer INDELs are under stronger selection.

Using the best-fitting DFE (Table 2.5), the proportion of INDEL substitutions in the CDS regions fixed by positive selection in the *D. melanogaster* lineage, α, is 83.7% (100% for insertions and 81.8% for deletions). These α estimates are comparable to previous estimates for SNP substitutions in CDS regions (Andolfatto *et al.*, 2011; Schneider *et al.*, 2011).

As mentioned above, some non-coding INDELs are probably non-neutral, as suggested by the negative Tajima's D value (Table 2.4). Our use of these variants as the neutral reference are for several practical reasons. Although using INDELs in "dead-on-arrival" transposable elements as neutral reference may be preferable (Petrov, 2002a), calling variants from repetitive regions using short-read data is highly prone to error (Li, 2014). Using data from the 8-30bp region of short introns $\leq$ 65bp, which are also putatively neutral (Parsch *et al.*, 2010), is also problematic because of evidence for selection maintaining intron size (Leushkin *et al.*, 2013; Parsch, 2003; Ptak and Petrov, 2002). Note that Tajima's D is more negative for INDELs in CDS regions than for those in non-coding regions, suggesting that the latter are probably under weaker purifying selection (Table 2.4). If this is the case, our method tends to underestimate the strength of purifying selection on INDELs in CDS regions, as suggested by the simulation results presented in supplementary Table A.2. This should lead to an overestimation of $\bar{\mu}$, the average fixation rate (Eq. (2.18)), which should in turn put a downward pressure on the estimation of α (Eq. (2.19)). However, biases in α also depend on the way selection on non-coding INDELs alters divergence. For example, if fixations of beneficial non-coding INDELs are so common that d_S is greater than the divergence level expected under neutral evolution, then this combined with the overestimation of $\bar{\mu}$ can lead to a substantial underestimation of α. In contrast, if most non-coding INDELs are selected against and d_S is much smaller than the neutral expectation, it may offset the effect caused by the overestimation of $\bar{\mu}$ and result in an overestimation of α.

2.4.2.3 Inferring the DFE and α using 4-fold degenerate sites as the neutral reference

To check the robustness of our results, we conducted a second set of analyses without using non-coding INDELs. We extended our model such that it can infer the DFE for INDELs in CDS regions using 4-fold sites as the neutral reference. We chose 4-fold sites instead of the 8-30bp region of short introns $\leq$ 65bp because 4-fold sites are probably not under ongoing selection on codon usage in *D. melanogaster*, and are similar to short introns in multiple aspects of polymorphism patterns (Jackson *et al.*, 2017). Considering the parameter richness of the models, using 4-fold SNPs as the neutral reference should help statistical inference because they are much more numerous than short-intron SNPs.

We used the following approach to obtain neutral divergence for INDELs along the *D. melanogaster* lineage. The nucleotide divergence in the 8-30bp region of short introns $\leq$ 65bp is 0.0674 (B. Jackson personal communication). In a mutation accumulation experiment (Schrider *et al.*, 2013), it was found that the rate to point mutations is 12.2 times higher than that to short INDELs, and that the rate to deletions is 5 times higher than that to insertions (averaging across the two genetic backgrounds considered therein). Thus, an estimate of neutral INDEL divergence can be obtained as $0.0674/12.2 = 0.0055$, and the corresponding estimates for insertions and deletions are 9.2×10^{-4} and 0.0046, respectively.

Due to the use of 4-fold sites as the neutral reference, it is no longer appropriate to assume that the mutation rate is the same between the selected and neutral regions. Given the evidence that the DFE for INDELs probably features a class of strongly deleterious mutations that make little contribution to polymorphism, allowing the selected and neutral regions to have their separate mutation rates is likely to cause the model to underestimate both the mutation rate in the selected region and strength of purifying selection, as confirmed by simulation results presented in supplementary Table A.3. An underestimation of the strength of purifying selection is likely to cause an underestimation of α. We observed this in our dataset – α for all INDELs obtained from the best-fitting DFE for this analysis (supplementary Table A.4) is only 21.7%, much smaller than the value of 83.7% when non-coding INDELs were used as the neutral reference (Table 2.5).

To resolve the above problem, we again made use of the information reported in the aforementioned mutation accumulation experiment (Schrider *et al.*, 2013). Specifically, we further extended our model, so that the mutation rate ratio between SNPs and INDELs, and that between deletions and insertions, were fixed at 12.2 and 5, respectively. As shown in Table 2.5 (see also supplementary Table A.5), the best-fitting DFE has $C = 2$ class of sites, with one under weak selection, and the other being strongly deleterious. The α estimates for all INDELs, insertions and deletions are, respectively, 71.5%, 59.7%, and 81.3%.

To make sure that the above results are not dependent on our use of the mutation rate ratios estimated by Schrider *et al.* (2013), we repeated the analysis using ratios obtained by either Petrov and Hartl (1998) (SNP/INDEL = 6.9 and deletion/insertion = 8.7) or Haag-Liautard *et al.* (2007) (SNP/INDEL = 4.2 and deletion/insertion = 3.0) (supplementary Table A.6). In both cases, the best-fitting DFE has $C = 2$ classes of selected sites, under weak and strong selection, respectively (supplementary Tables A.7 and A.8). Furthermore, estimates of the strength of purifying selection acting on sites in the weakly selected class are almost identical regardless of the choice of mutation rate ratios (supplementary Table A.9). Thus, unsurprisingly, all three analyses also produce very similar α estimates (supplementary Table A.9). Overall, these results are consistent with those based on non-coding INDELs and suggest that a substantial fraction of INDEL substitutions were fixed by positive selection.

2.5 Materials and Methods

2.5.1 Numerical details

We used numerical routines provided by the GNU Scientific Library (GSL; https://www.gnu.org/software/gsl/) to perform the integration in (2.3) numerically. For the continuous model (e.g., (2.9)), the integral was evaluated using Gaussian quadrature, which was implemented based on a routine included in the R package `statmod` (https://cran.r-project.org/web/packages/statmod/index.html). Maximum likelihood estimates of the model parameters were obtained by both gradient-based and derivative-free optimization algorithms implemented in the `NLopt` package (http://ab-initio.mit.edu/wiki/index.php/NLopt). To ensure the global maximum was found, we initialised the search algorithm using multiple randomly selected starting points.

2.5.2 Simulations

We performed parameter estimation using our program, `anavar`, on random samples simulated using Mathematica (http://www.wolfram.com/). Because the generation of simulated data is separate from the numerical routines we used to implement `anavar`, this set-up can help verify the numerical robustness of `anavar`. Note that, in all simulations, we only used the models to analyse variants from selected regions because we wanted to find out how much information we could obtain by analysing them alone. Including neutral variants, as routinely done in real data analysis, may help to increase the accuracy of parameter estimation. So our choice should give us a rather conservative assessment of the methods' performance.

In addition to testing whether the data contained enough information for all the parameters to be estimated, we also assessed how well a model could predict the average fixation rate, $\bar{\mu}$ (expressed in units of $2N_e$ generations). As an example, if nonsynonymous polymorphism data are fitted to the discrete SNP model, $\bar{\mu}$ can be estimated as

$$\bar{\mu} = \frac{1}{\hat{\theta}} \sum_{c=1}^{C} \frac{\hat{\theta}_c \hat{\gamma}_c}{1 - e^{-\hat{\gamma}_c}} \tag{2.18}$$

where $\hat{Z}$ signifies the MLE of parameter Z and θ is defined by (2.1). Understanding the ability to accurately estimate $\bar{\mu}$ is important because it is needed for estimating α, the proportion of substitutions fixed by positive selection, which can be written as,

$$\alpha = \frac{d_N - d_S \bar{\mu}}{d_N} \tag{2.19}$$

where d_N and d_S are the numbers of selected (e.g., nonsynonymous) and neutral (e.g., synonymous) substitutions per site, respectively (Eyre-Walker and Keightley, 2009).

We did not generate simulated data from models with demographic changes and selection at linked sites because the effectiveness of the method of Eyre-Walker *et al.* (2006) in controlling for these confounding factors have been studied extensively (Eyre-Walker *et al.*, 2006; Galtier, 2016; Glémin *et al.*, 2015; Jackson *et al.*, 2017; Muyle *et al.*, 2011; Tataru *et al.*, 2017).

2.5.3 The *Drosophila melanogaster* dataset

This dataset consisted of 17 Rwandan individuals as described in Jackson *et al.* (2015, 2017) and made available by the *Drosophila* Population Genomics Project (Pool *et al.*, 2012).

2.5.3.1 Variant calling

INDEL realigned BAM files were obtained from Jackson *et al.* (2017). Initial genotype calling was performed with the HaplotypeCaller and GenotypeGVCF (with the `-includeNonVariantSites` flag to output genotype calls at both variant and non-variant positions) tools from GATK 3.7 (DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). Variant quality score recalibration (VQSR) requires one 'truth set' for SNPs and one for INDELs. To generate the truth sets, we intersected the raw variants called from GATK with variants called from SAMtools (version 1.2) (Li *et al.*, 2009). The consensus data was further filtered using the GATK best practice hard filters (for SNPs: QD $<$ 2.0, MQ $<$ 40.0, FS $>$ 60.0, SOR $>$ 3.0, MQRankSum $<$ -12.5, ReadPosRankSum $<$ -8.0; for INDELs: QD $<$ 2.0, ReadPosRankSum $<$ -20.0, FS $>$ 200.0, SOR $>$ 10.0; see `https://software.broadinstitute.org/gatk/guide/article?id=3225`). Variants with coverage more than twice, or less than half, the mean coverage of 20X were excluded, along with variants falling into regions identified by `RepeatMasker` (`http://www.repeatmasker.org`). Multiallelic sites were excluded along with SNPs falling within INDELs and INDELs greater than 50bp. We ran VQSR separately for

SNPs and INDELs, retaining variants that fell within the 95% tranche cut-off as in Jackson *et al.* (2017). The passing variants were then re-filtered as above with the exception of the GATK hard filters which were not reapplied.

2.5.3.2 Multi-species alignments and polarisation

Multi-species alignments were generated between *D. melanogaster* (v5.34), *D. simulans* (Hu *et al.*, 2013) and *D. yakuba* (v1.3) using *D. melanogaster* as reference. Firstly pairwise alignments were created using LASTZ (Harris, 2007). These were then chained and netted using axtChain and chainNet, respectively (Kent *et al.*, 2003). Single coverage was ensured for the reference genome using single_cov2.v11 from the MULTIZ package (Blanchette *et al.*, 2004) and the pairwise alignments were aligned with MULTIZ.

Variants were polarised using the whole genome multi-species alignment and a parsimony approach, where either the alternate or the reference allele had to be supported by all outgroups in the the alignment to be considered ancestral. The site-frequency spectra for insertions and deletions in different genomic regions are presented in supplementary Figure A.4.

2.5.3.3 Annotation

Variants were annotated as either intronic, intergenic or CDS using the *D. melanogaster* GFF annotation file (version 5.34, available from: `ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.34_FB2011_02/gff/`). Fourfold degenerate and zerofold degenerate SNPs in CDS regions were annotated using coordinates obtained from the *D. melanogaster* CDS fasta sequences (version 5.34, available from: `ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.34_FB2011_02/fasta/dmel-all-CDS-r5.34.fasta.gz`).

2.5.3.4 Summary statistics

Nucleotide diversity (π) (Tajima, 1983), Watterson's θ (Watterson, 1975) and Tajima's D (Tajima, 1989) were calculated for variants in non-coding (intronic and intergenic) and coding regions, as well as for 0-fold and 4-fold degenerate SNPs. The numbers of callable sites used to obtain per-site estimates was taken to be the number of sites in each region that were called in the "all sites" VCF file and passed the filters described previously. Additionally for polarised variants the number of callable sites was reduced to those that could be polarised by our parsimony approach.

To obtain rates of divergence at nonsynonymous and synonymous sites, denoted by d_N and d_S, CDS regions were extracted from the multi-species alignment using the coordinates from the *D. melanogaster* CDS fasta alignment file. CDS alignments were removed if they were not in frame, did not start with a start codon, did not end with a stop codon or contained premature stop codons. Additionally any codons with missing data were dropped. For each gene we retained only the longest transcript. This data was then analysed using codeml in PAML (Yang, 2007) with a one ratio model to obtain d_N and d_S.

2.6 Supplementary Material

The new models have been implemented in a user-friendly package `anavar`, which is freely available at `http://zeng-lab.group.shef.ac.uk`. In addition to the models developed herein, `anavar` also contains implementations of several other widely-used models for estimating the DFE (i.e., Eyre-Walker *et al.*, 2006) and for studying GC-biased gene conversion (gBGC) (i.e., Glémin *et al.*, 2015). All scripts used for the `anavar` simulation analyses are available at `https://github.com/henryjuho/anavar_simulations`. Additionally, all scripts used in the *D. melanogaster* analyses can be found at `https://github.com/henryjuho/drosophila_indels`.

2.7 Acknowledgments

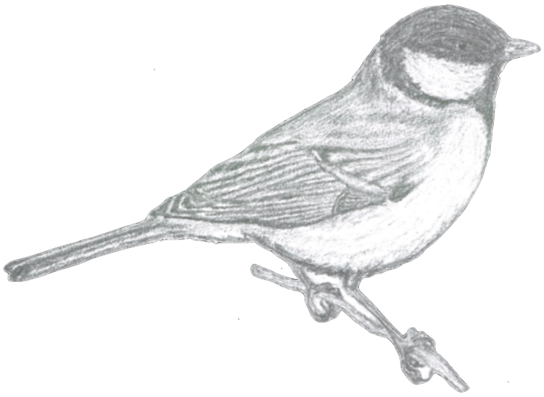
We thank Ben Jackson and Jose Campos for providing valuable assistance in the analysis of the *Drosophila* data. This work was supported by a PhD studentship funded by the Department of Animal and Plant Sciences, University of Sheffield, to H.J.B. Support was also provided by the Natural Environment Research Council via a research grant awarded to K.Z. (NE/L005328/1).

Chapter 3

The impact of natural selection on short insertion and deletion variation in the great tit genome

Authors: **Henry J. Barton** and **Kai Zeng**

This chapter has been published in the *Genome Biology and Evolution* under the same title, see Barton and Zeng (2019). It is shown here in its published form with additional formatting changes.



3.1 Abstract

Insertions and deletions (INDELs) remain understudied, despite being the most common form of genetic variation after single nucleotide polymorphisms. This stems partly from the challenge of correctly identifying the ancestral state of an INDEL and thus identifying it as an insertion or a deletion. Erroneously assigned ancestral states can skew the site frequency spectrum, leading to artificial signals of selection. Consequently, the selective pressures acting on INDELs are, at present, poorly resolved. To tackle this issue, we have recently published a maximum likelihood approach to estimate the mutation rate and the distribution of fitness effects (DFE) for insertions and deletions. Our approach estimates and controls for the rate of ancestral state misidentification, overcoming issues plaguing previous INDEL studies. Here we apply the method to INDEL polymorphism data from 10 high coverage ($\sim 44X$) European great tit (*Parus major*) genomes. We demonstrate that coding INDELs are under strong purifying selection with a small proportion making it into the population ($\sim 4\%$). However, among fixed coding INDELs, 71% of insertions and 86% of deletions are fixed by positive selection. In non-coding regions we estimate $\sim 80\%$ of insertions and $\sim 52\%$ of deletions are effectively neutral, the remainder show signatures of purifying selection. Additionally, we see evidence of linked selection reducing INDEL diversity below background levels, both in proximity to exons and in areas of low recombination.

3.2 Introduction

Insertion and deletion mutations (INDELs) are an important source of genetic variation, often separated into long and short INDELs due to different calling approaches required for longer variants. There is one short INDEL (here ≤50bp) for every 8 single nucleotide polymorphisms (SNPs) in humans (Montgomery *et al.*, 2013), representing a significant proportion of variation. Short INDELs have been implicated in a range of genomic evolutionary processes, such as the evolution of genome size (Hu *et al.*, 2011; Nam and Ellegren, 2012; Petrov, 2002b; Sun *et al.*, 2012). INDELs arguably contribute more to sequence divergence, in terms of the number of base differences, than SNPs (Britten, 2002). Additionally it has been suggested that short INDELs may be instrumental in maintaining an optimal intron size (Parsch, 2003; Presgraves, 2006).

INDEL studies, however, are under-represented in the literature. In part, this is due to the need to categorise INDELs into insertions and deletions, which requires knowledge of the ancestral state for each variant. This can be obtained using multi-species genome alignments. However, INDELs disproportionately occur in repetitive sequence contexts (Ananda *et al.*, 2013; Montgomery *et al.*, 2013), which are notoriously problematic to align (Earl *et al.*, 2014). Where alignments are successful they are hampered by high rates of ancestral allele misidentification, due to homoplasy. The result is a proportion of deletions are mistakenly identified as insertions (and vice versa), which can confound estimates of selection (Kvikstad and Duret, 2014) (see figure 2.1).

Despite the difficulty of analysing INDEL data, a number of characteristics have been widely reported for INDELs. INDEL mutation is consistently biased towards deletions across a diverse range of organisms (Hu *et al.*, 2011; Keightley *et al.*, 2009; Kvikstad and Duret, 2014; Nam and Ellegren, 2012; Presgraves, 2006; Taylor *et al.*, 2004). Additionally, polymerase slippage has emerged as the predominant force driving short INDEL generation, explaining $\sim 75\%$ of events in repetitive hotspot regions (Montgomery *et al.*, 2013) and $\sim 50\%$ of events in non-hotspot regions (Montgomery *et al.*, 2013; Taylor *et al.*, 2004).

In terms of the selective pressures acting on INDELs, deletions consistently segregate at lower frequencies than insertions, both in genes (Sjödin *et al.*, 2010) and genome-wide (Chintalapati *et al.*, 2017), which has been interpreted as stronger purifying selection acting on deletions. A mechanistic explanation is that deletions have two breakpoints relative to an insertion's one, so are more likely to hit an important motif (Petrov, 2002b; Sjödin *et al.*, 2010). The difference in mean allele frequencies of the two types of variation has also been explained as selection acting on insertions (Ometto *et al.*, 2005). Concordantly, a number of studies have inferred elevated fixation rates for insertions from comparisons of the ratio of deletion to insertion events (rDI) between polymorphism data and divergence data (Chintalapati *et al.*, 2017; Leushkin and Bazykin, 2013; Presgraves, 2006; Sjödin *et al.*, 2010). This fixation bias is in line with a number explanations such as selection on insertions to maintain intron lengths (Ometto *et al.*, 2005; Parsch, 2003; Presgraves, 2006) or insertion biased gene conversion (Leushkin and Bazykin, 2013). However, Kvikstad and Duret (2014) demonstrate the existence of mutation hotspots in repetitive regions, and cryptic hotspots in non-repetitive regions, which could explain the fixation biases by elevating rates of ancestral state misidentification. They also show that differences in the rate of ancestral misidentification between polymorphism data and divergence data make McDonald-Krietman type tests (McDonald and Kreitman, 1991), which in an INDEL context compare polymorphic and fixed numbers of deletions and insertions (for example see Chintalapati *et al.*, 2017), particularly prone to false signatures of fixation bias.

Avian genomes provide a good system for working on INDELs, thanks to their markedly conserved karyotypes and synteny, characterised by having few large macro-chromosomes and many smaller micro-chromosomes (Hansson *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014; Zhang *et al.*, 2014). Not only does this facilitate genome alignments for ancestral state identification, but obligate crossing over elevates recombination rates on micro-chromosomes, driving large intra-genomic variation in recombination (Backström *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014). This provides power for associating diversity levels with recombination rates. As a result, birds have been the focus of a number of INDEL studies. Nam and Ellegren (2012) propose that high recombination rates drive elevated small deletion rates on micro-chromosomes and might have caused genome contraction along the lineage leading to birds. Additionally, Rao *et al.* (2010) show a positive correlation between INDEL density and recombination rate in chicken (*Gallus gallus*) introns. Whilst this may suggest the impact of linked selection, the use of unpolarised INDEL data means it cannot be distinguished from the impact of a recombination driven mutational bias, such as proposed by Nam and Ellegren (2012). Furthermore, previous work has been constrained by utilising partial sequencing approaches and neutral markers, negating the formation of a genome wide picture of INDEL diversity (Brandstrom and Ellegren, 2007; Nam and Ellegren, 2012; Rao *et al.*, 2010). Thus, despite the advantages of an avian system, the role of natural selection in shaping INDEL diversity in birds is poorly resolved.

Most existing work looking at selection on INDELs has relied upon approaches susceptible to the confounding effects of ancestral state misidentification. There also has been little effort to directly infer unbiased selection coefficients for INDELs, in different genomic contexts. To bridge this gap we recently published our maximum likelihood model 'anavar' for estimating the mutational and selective parameters for INDELs, whilst simultaneously estimating and controlling for ancestral state misidentification and the confounding effects of demography (Barton and Zeng, 2018). Here, we apply this approach to INDEL polymorphism data from 10 European great tit (*Parus major*) genomes from Corcoran *et al.* (2017). We investigate the selective pressures acting on INDELs across the great tit genome and estimate selection coefficients and the proportion of substitutions fixed by positive selection (α) in coding regions. We also seek to address how INDEL diversity changes with distance from coding regions and assess the impact of linked selection on INDEL variation, an area understudied in the literature so far. The great tit genome is particularly well positioned to address these questions with an abundance of current genomic resources available including a well annotated reference genome, high coverage resequencing data, and replicated linkage maps (Corcoran *et al.*, 2017; Laine *et al.*, 2016; van Oers *et al.*, 2014).

3.3 Materials and Methods

3.3.1 The great tit dataset

The great tit dataset consisted of 10 European males (1280, 1485, 15, 167, 249-R, 318, 61, 917, 943-R and TR43666) from a subset of sampling locations in Laine *et al.* (2016) as described in Corcoran *et al.* (2017). The mean coverage of the sample is 44X.

3.3.2 Data preparation and variant calling

Base quality score recalibrated and INDEL realigned BAM files, and an all-sites VCF file containing raw variant calls produced by GATK (version 3.4) (DePristo *et al.*, 2011; McKenna *et al.*, 2010; Van der Auwera *et al.*, 2013) were obtained from Corcoran et al. (2017).

Variant quality score recalibration (VQSR) was then performed for INDELs. This step requires a set of high confidence variants. To generate this data set, we intersected the raw variants called from GATK with variants called with SAMtools (version 1.2) (Li *et al.*, 2009). The resulting variants were filtered using the GATK best practice hard filters ($QD < 2.0$, $ReadPosRankSum < -20.0$, $FS > 200.0$, see `https://software.broadinstitute.org/gatk/guide/article?id=3225`; last accessed October 1, 2018). Variants with coverage more than twice, or less than half, the mean coverage of 44X were excluded, along with variants falling in repeat regions identified by RepeatMasker (Smit *et al.*, 2013). INDELs with more than two alleles of different length (multiallelic sites) were excluded and INDELs greater than 50bp. Post VQSR, we retained variants that fell within the 99% tranche cut-off. The passing variants were then re-filtered as above with the exception of the GATK hard filters, which were not reapplied.

For SNPs, variants passing the 99% tranche cut-off in the data set of Corcoran *et al.* (2017) were obtained and subject to the same post VQSR hard filters as described above for INDELs.

3.3.3 Multispecies alignment and polarisation

We created a multispecies alignment between zebra finch (*Taeniopygia guttata*) (Warren *et al.*, 2010) (version: TaeGut3.2.4, available from: `ftp://ftp.ensembl.org/pub/release-84/fasta/taeniopygia_guttata/dna/`; last accessed October 1, 2018), flycatcher (*Ficedula albicollis*) (Ellegren *et al.*, 2012) (version: FicAlb1.5, available from: `http://www.ncbi.nlm.nih.gov/genome/?term=flycatcher`; last accessed October 1,

2018) and great tit (version 1.04) (Laine *et al.*, 2016) with the MULTIZ package (Blanchette *et al.*, 2004) per chromosome, following the pipeline described in Corcoran *et al.* (2017).

The ancestral states of each variant were then inferred using a parsimony approach where all out-groups were required to match either the reference, or the alternate, allele in the great tit in order to assign it as ancestral.

3.3.4 Variant annotation

All variants were annotated as coding, intronic or intergenic using the great tit annotation (version 1.03) (available from: `ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/522/545/GCF_001522545.1_Parus_major1.0.3/GCF_001522545.1_Parus_major1.0.3_genomic.gff.gz`; last accessed October 1, 2018). Additionally the possible locations of fourfold degenerate sites, zerofold degenerate sites and nonsense mutations were identified using the great tit coding sequence fasta file (version 1.03) (available from: `ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/522/545/GCF_001522545.1_Parus_major1.0.3/GCF_001522545.1_Parus_major1.0.3_cds_from_genomic.fna.gz`; last accessed October 1, 2018) SNPs at these positions were then identified.

We identified ancestral repeats (specifically, LINEs) by intersecting the RepeatMasker coordinates for each species with our whole genome alignment and identifying positions annotated as LINEs in all three species. Variants within these regions were identified from the VCF files prior to filtering and were then filtered as described previously, with the exception of the repeat filtering.

We identified callable sites for use in the calculation of summary statistics and our anavar analyses by applying our filters to the original all-sites VCF file and restricting the sites to those that we could polarise.

3.3.5 Summary statistics

We calculated nucleotide diversity (π) (Tajima, 1983) and Tajima's D (Tajima, 1989) for INDELs and SNPs both genome-wide and in ancestral repeats (ARs), introns, intergenic regions and coding sequences (CDS). In coding regions we analysed mutations that preserve the reading frame (in-frame: SNPs, and INDELs a multiple of three in length) and those that shift the reading frame (frame-shift: remaining INDELs) separately. For SNPs we also calculated these statistics for fourfold degenerate sites, zerofold degenerate sites and nonsense mutations. Additionally, we calculated Tajma's D for each INDEL length group separately. Note that while classically π refers to the average number of nucleotide differences (Tajima, 1983), for INDELs we are measuring the average number of mutation differences without accounting for the number of bases a given INDEL encompasses.

We also calculated Tajima's D and π using the site frequency spectrum corrected for orientation errors. We took the model estimates of polarisation error for the regions under consideration (see table B.1), and solved the system of linear equations:

$$\phi_i^{ins,obs} = (1-\epsilon^{ins})\phi_i^{ins} + \epsilon^{del}\phi_{n-i}^{del} \tag{3.1}$$

$$\phi_{n-i}^{del,obs} = (1-\epsilon^{del})\phi_{n-i}^{del} + \epsilon^{ins}\phi_i^{ins} \tag{3.2}$$

for $1 \leq i < n$, where $\phi_i^{ins,obs}$ ($\phi_i^{del,obs}$) is the observed number of insertions (deletion) of frequency i, ϵ^{ins} (ϵ^{del}) the probability that the ancestral state of an insertion (deletion) is incorrectly identified, and ϕ_i^{ins} (ϕ_i^{del}) the underlying (unobserved) site frequency spectrum for insertions and deletions. Tajima's D and π were then calculated using ϕ_i^{ins} and ϕ_i^{del}.

We calculated the distribution of INDEL lengths from our VCF file, both genome-wide and in CDS regions. Within CDS regions we calculated the proportion of in-frame

INDELs per gene. We calculated this proportion both for all genes and for a set of conserved genes identified in Corcoran *et al.* (2017).

Divergence estimates for INDELs were calculated by counting the number of fixation events unique to the great tit lineage in our whole genome alignment, and dividing by the number of sites that were aligned in all three species for each region analysed (CDS, AR, intron and intergenic). For SNPs we created concatenated FASTA files for each region (CDS, AR, intron and intergenic), and obtained a pairwise distance matrix using APE (Paradis *et al.*, 2004) in R (R Core Team, 2015). The pairwise distance estimates were then used to get an estimate for the branch leading to the great tit.

3.3.6 DFE analysis

To estimate the distribution of fitness effects (DFE) for insertions and deletions we used the "neutralINDEL_vs_selectedINDEL" model in the anavar package (Barton and Zeng, 2018) (available from: `http://zeng-lab.group.shef.ac.uk/wordpress/?page_id=28`; last accessed October 1, 2018). The package controls for the confounding effects of polarisation error and demography (Barton and Zeng, 2018). We fitted two types of models for the DFE. The first type fits a discrete number of site classes (c) to the data, each class having its own scaled selection coefficient, $\gamma = 4N_e s$. The per-site scaled mutation rate, $\theta = 4N_e\mu$, may be equal across sites (the equal mutation rate model), or be different between the neutral sites and the focal sites (the variable mutation rate model). Finally, the model has polarisation error parameters, ϵ^{ins} and ϵ^{del}, for both insertions and deletions. The second type of model is similar, but assumes continuous gamma distributions for the selection coefficients for insertions and deletions. Different variants of these two types of model were fitted (e.g., with different numbers of site classes and with the mutation rate being either equal or variable) and were compared using Akaike information criterion (AIC).

We used INDELs in ancestral repeats (as described previously) as neutral reference, and applied the models separately to CDS INDEL data and to non-coding INDEL data. For coding sequence data we assumed the equal mutation rate model. This is necessary in order to estimate the proportion of substitutions fixed by positive selection (α), as well as estimating the proportion of strongly deleterious variants that do not contribute to polymorphism. We calculated α using equation 19 from Barton and Zeng (2018). For non-coding data we employed the variable mutation rate model, which fitted the data better than the equal mutation rate model. We will explore the effects of model choice on our results in the Discussion.

3.3.7 Exon proximity analysis

To investigate the impact of linked selection on INDEL diversity patterns in regions adjacent to coding sequences we extracted INDELs and numbers of callable sites in 2kb adjacent windows moving away from exons up to a maximum distance of 100kb. The data from all windows at each distance was then binned, creating 50 distance bins. We ran each of the resulting datasets through the anavar package. We fitted the "neutralINDEL_vs_selectedINDEL" model with a continuous γ distribution and variable mutation rates, as this was the best fitting model for non-coding INDELs (table B.4). We used the same neutral reference as in our previous analysis. The relationship between the model's θ estimates and distance from exons was tested with Spearman's correlations using the 'cor.test' function in R (R Core Team, 2015). We repeated this analysis using π estimates for insertions and deletions instead of the model's mutation rate estimates.

To look at the relative contributions of different selective site classes to INDEL diversity in each window, we separated our θ estimates into θ for sites with $0 \leq \gamma \leq 1$ and θ for $\gamma > 1$ using the model outputs, we repeated the correlation analysis for these datasets.

To assess to what extent the relationship between distance from exon and diversity was driven by bins close to exons, we generated downsized datasets by progressively removing bins, starting by removing the nearest bin, and then the next nearest, and so on, up until only the furthest two bins were left. We reported the Spearman's correlation coefficient (ρ) and the significance for each down-sampled dataset.

3.3.8 Recombination correlation analysis

To investigate the relationship between local recombination rate and the action of linked selection we divided the great tit genome into 2Mb non-overlapping windows. We extracted non-coding INDEL calls for each window from our VCF file, excluding windows with less than 500 polarisable INDELs. As we lacked sufficient data to obtain a regional neutral reference for each window, we were unable to apply our model based approach. Instead we calculate π and Tajima's D for each window. We also estimated non-coding INDEL divergence per window as described previously.

Mean recombination rate was estimated per window. This was achieved by estimating a point recombination rate for every INDEL in the window, along with positions 2kb up and down stream of each variant and taking a mean across all these values. The site specific recombination rates were estimated using the pipeline described in Corcoran *et al.* (2017). Briefly, we fitted 3rd order polynomials as a function of physical position versus map length for each chromosome using the great tit linkage map data (van Oers *et al.*, 2014). The derivative of each chromosome's polynomial was then used to estimate recombination rate at a given genomic position.

The relationships of Tajima's D and π with local recombination rate were analysed with Spearman's correlations using the 'cor.test' function in R (R Core Team, 2015). The relationship between π and recombination rate was also analysed using partial Spearman's correlations, with divergence estimates as a confounding variable, to control for the mutagenic effect of recombination, using the 'ppcor' package (Kim, 2015) in R.

3.3.9 Data Availability

Detailed documentation of the analysis pipeline along with all scripts used is available at `https://github.com/henryjuho/parus_indel` (last accessed October 1, 2018). The python scripts make use of the pysam python package (`https://github.com/pysam-developers/pysam`; last accessed October 1, 2018) and the anavar_utils package (`https://henryjuho.github.io/anavar_utils/`; last accessed October 1, 2018).

Table 3.1: Nucleotide diversity (π) for SNPs, INDELs (unpolarised), insertions (ins) and deletions (del) in different genomic contexts. Estimates in brackets corrected for polarisation error.

Context	π	π_{indel}	π_{ins}	π_{del}
Genome	0.00310	0.000356	0.000113 (0.000112)	0.000142 (0.000144)
ARs	0.00432	0.000363	0.000117 (0.000119)	0.000175 (0.000177)
Intergenic	0.00333	0.000378	0.000121 (0.000119)	0.000154 (0.000157)
Introns	0.00306	0.000361	0.000116 (0.000115)	0.000143 (0.000145)
CDS	0.00145	1.87×10^{-5}	3.61×10^{-6} (4.36×10^{-6})	5.25×10^{-6} (5.09×10^{-6})
In-frame	-	9.43×10^{-6}	1.71×10^{-6} (1.86×10^{-6})	3.00×10^{-6} (3.04×10^{-6})
Frame-shift	-	9.28×10^{-6}	1.90×10^{-6} (2.17×10^{-6})	2.24×10^{-6} (2.27×10^{-6})
4-fold	0.00369	-	-	-
0-fold	0.000586	-	-	-
Nonsense	2.45×10^{-5}	-	-	-

3.4 Results

3.4.1 Summary of the dataset

Using the high coverage resequencing data from Corcoran *et al.* (2017) we called polymorphic INDELs and SNPs according to a GATK based pipeline (Van der Auwera *et al.*, 2013). We polarised variants using a custom multi-species genome alignment and a parsimony based approach. Application of our data calling pipeline to the 10 European great tit samples yielded 10,259,689 SNPs and 1,162,517 short INDELs ($\leq$ 50bp), of which we could polarise 254,040 insertions and 329,506 deletions. This reduction in variants in the polarised dataset is mainly a result of gaps in the whole genome alignment and ‘hotspots’ where the INDEL breakpoints differ between species in the alignment (figure B.1).

Genome-wide diversity (π) for INDELs is around tenfold lower than that for SNPs. This scale of difference between the two forms of variation was found in all genomic regions analysed other than in CDS regions where INDEL diversity is close to 80 times lower than SNP diversity. Additionally, we see that within INDELs π is biased towards deletions in all regions (table 3.1).

When considering INDEL sequence length we observe that the length distribution is enriched in shorter variants, with 80% of INDELs less than 5bp long. Additionally, within coding sequences (CDS) we note that the length distribution is enriched in variants that

Table 3.2: Maximum likelihood parameter estimates for the best-fitting models for INDELs in CDS regions and non-coding regions. C defines the number of site class, θ the population scaled mutation rate, γ the population scaled selection coefficient, ϵ the polarisation error and α the proportion of INDEL substitutions driven by positive selection. Note where γ values are presented for the continuous model these are mean γ estimates and the product of the scale and shape parameters.

Model and DFE	Type	C	θ	γ	scale	shape	ϵ	α
CDS: Equal θ	INS	1	4.92×10^{-6}	-1.14	-	-	0.0799	
Discrete $C = 2$	INS	2	0.000134	-801	-	-	0.000307	71%
AR reference	DEL	1	8.32×10^{-6}	-2.70	-	-	0.0368	
	DEL	2	0.000206	-649	-	-	3.12×10^{-7}	86%
CDS: Equal θ	INS	1	4.79×10^{-6}	-0.264	-	-	0.0729	
Discrete $C = 2$	INS	2	0.000156	-897	-	-	0.000526	63%
NC reference	DEL	1	7.79×10^{-6}	-1.70	-	-	0.0366	
	DEL	2	0.000205	-629	-	-	0.00587	79%
Non-coding: Free θ	INS	-	0.000170	-53.6	1553	0.0345	0.0110	-
Continuous	DEL	-	0.000293	-75.5	715	0.106	0.0166	-

are a multiple of three in length, in other words, mutations that preserve the reading frame (in-frame) (figure B.2). This enrichment is even more pronounced in conserved genes (figure B.3). To further investigate the differences between in-frame and frame-shifting INDELs, we first note that it is far more likely for an INDEL mutation to have a length that is not a multiple of three than otherwise. This can be seen by the fact that, in putatively neutrally evolving ancestral repeat (AR) regions, π values for insertions and deletions with lengths not a multiple of three are 9.8×10^{-5} and 1.4×10^{-4} respectively, whereas for those with lengths a multiple of three, the values are 1.9×10^{-5} and 3.4×10^{-5}. When we consider this in terms of the ratio of AR to CDS diversity (using the CDS π values in table 3.1), for mutations that shift the reading frame we get a ratio of 52 for insertions and 63 for deletions, whereas for in-frame mutations the ratios are both 11. This indicates a much larger reduction in diversity for frame-shifting INDELs, and this reduction is more pronounced for deletions, supporting the idea that they are more deleterious.

In general, ancestral repeats have the highest diversity level and the least negative Tajima's D for both INDELs and SNPs (table 3.1 and figure 3.1a). This supports our decision to use them as a putatively neutral reference in the subsequent analyses. The fact that Tajima's D values are consistently negative in AR regions (figure 3.1a) is consistent with a recent population expansion for the great tit, as previously reported

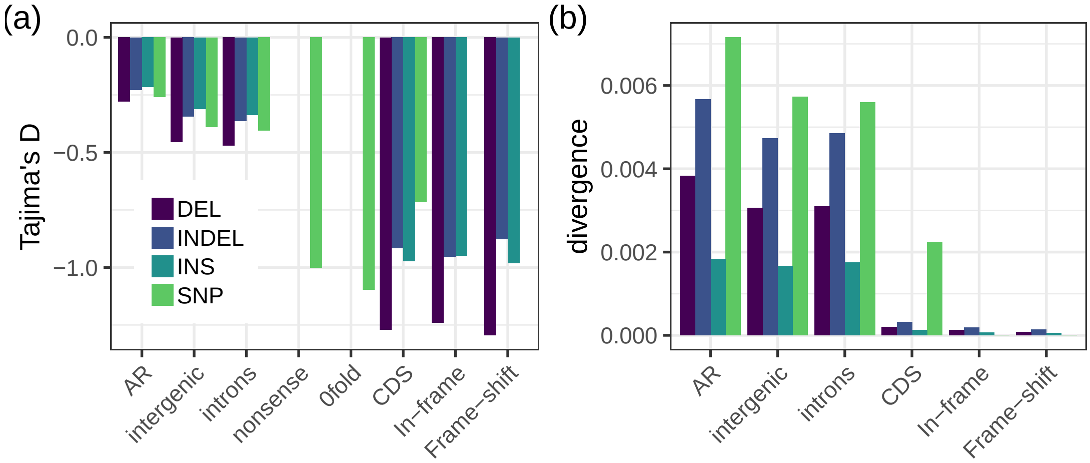


FIGURE 3.1: Tajima's D estimates for SNPs, INDELs (unpolarised), insertions (INS) and deletions (DEL) in different genomic contexts. Divergence estimates for SNPs are presented as the true divergence divided by 10.

(Corcoran *et al.*, 2017; Laine *et al.*, 2016). Intronic and intergenic regions have similar diversity patterns across all mutation types, so we grouped them as 'noncoding' in subsequent analyses. Tajima's D values for the unpolarised INDELs in CDS regions are similar to those for 0-fold SNPs and SNPs that cause premature stop codons (nonsense mutations). However when polarised, we see that deletions in CDS regions have the most negative Tajima's D of all (figure 3.1a). In non-coding regions, Tajima's D is negatively correlated with INDEL size for both insertions (Spearman's $\rho = -0.95$, $p < 2.2 \times 10^{-16}$) and deletions (Spearman's $\rho = -0.40$, $p = 0.0038$), suggesting that longer variants are probably more deleterious (figure B.4). In coding regions we lack power when sub-setting INDELs by length (figure B.4).

The patterns reported above are mirrored by the divergence estimates. The highest divergence is seen in ARs. Intergenic and intronic regions have similar divergence levels, and both have lower divergence than ARs. In CDS regions divergence is lowest, 14 times lower than the genome-wide average for INDELs. SNP divergence is around tenfold higher than INDEL divergence in non-coding regions, in line with π estimates. In CDS regions SNP divergence is seventyfold higher than INDEL divergence (figure 3.1b). These results are robust to polarisation error (table 3.1, figure B.5).

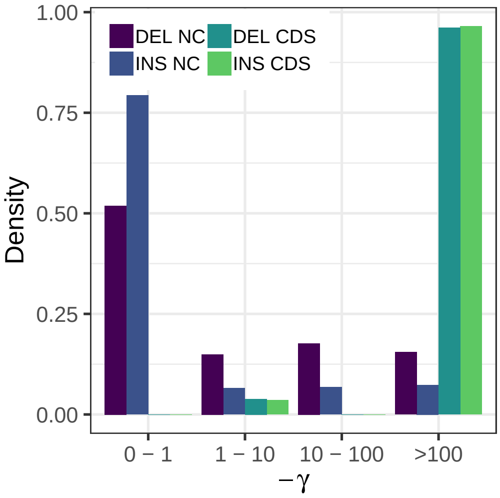


FIGURE 3.2: Distribution of fitness effects for non-coding insertions (INS NC), non-coding deletions (DEL NC), coding insertions (INS CDS) and coding deletions (DEL CDS), shown as the proportion of mutations falling into different selection coefficient (γ) bins.

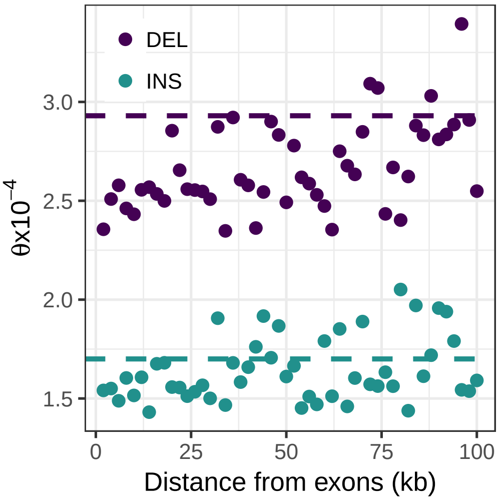


FIGURE 3.3: Relationship between mutation rate estimates (θ) for insertions (turquoise) and deletions (purple) and distance from exons in 2kb windows. Dashed lines represent the genome wide average mutation rate for non-coding variants, as show in table 3.2.

3.4.2 The distribution of fitness effects

To describe the distribution of fitness effects (DFE) for INDELs we fitted 4 distinct DFEs to coding and non-coding data separately. For coding data the model assumes equal mutation rates between neutral and focal sites, a requirement to calculate the proportion of substitutions fixed by positive selection (α). For non-coding data where α was not calculated, this assumption was relaxed and mutation rates were free to vary (see Materials and Methods 3.3). The best-fit model for each case is reported in table 3.2.

The best-fit INDEL DFE (according to AIC, see table B.2) in coding regions is bimodal, characterised by a class of strongly deleterious insertions and deletions making up 96% of sites and a class of weakly deleterious insertions and deletions for the remaining 4% of sites (figure 3.2). For those variants with weakly negative γ estimates (i.e. those segregating in our sample) deletions are more deleterious, however for the strongly deleterious class of INDELs insertions have the more negative selection coefficient. We subsequently estimate the proportion of INDEL substitutions fixed by positive selection (α) at 71%

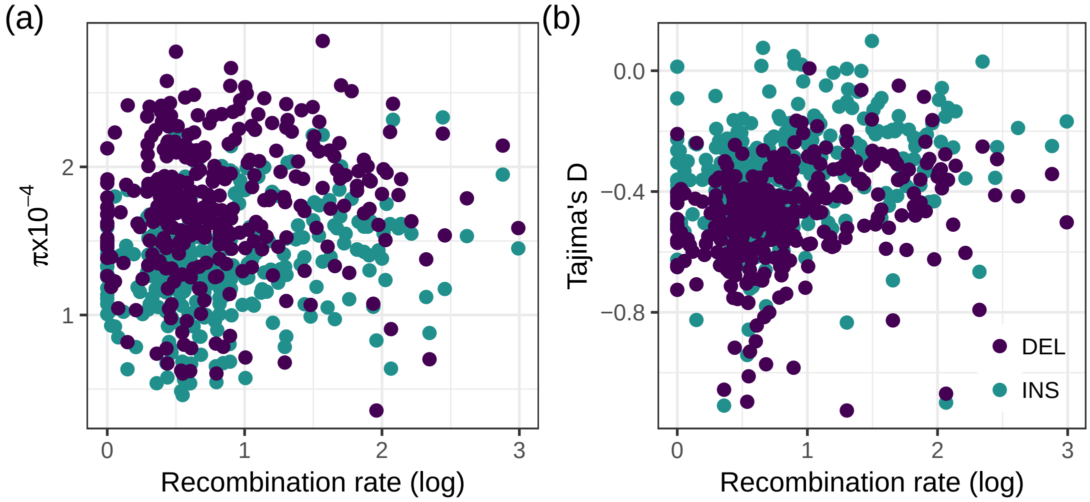


FIGURE 3.4: The relationship between local recombination rate (log transformed) and π (a) and Tajima's D (b) for both insertions (turquoise) and deletions (purple)

for insertions and 86% for deletions (table 3.2). When we run this analysis using a non-coding neutral reference we recapture a very similar bimodal DFE, but with slightly lower α values, 63% for insertions and 79% for deletions (table 3.2 and table B.3).

The non-coding INDEL data is best fit by a continuous gamma distribution of fitness effects (table B.4). We see small shape parameter estimates of 0.0345 for insertions and 0.106 for deletions (table 3.2), describing a DFE enriched in effectively neutral variants. When binning this gamma distribution into four $-\gamma$ categories ($0-1$, $1-10$, $10-100$ and > 100) we see that $\sim$ 80% of insertions and $\sim$ 52% of deletions in non-coding regions have γ estimates between 0 and -1 and can be considered as effectively neutral. The remaining proportions of variants are evenly distributed between the other 3 selective categories (figure 3.2). For non-coding and coding data there is a marked deletion bias with the deletion to insertion ratio (rDI) estimated at 1.5 in coding regions and 1.7 in non-coding regions.

3.4.3 The impact of linked selection

To test for evidence of linked selection acting on INDELs, we obtained estimates of the scaled insertion and deletion mutation rates (θ_{ins} and θ_{del} respectively) in 2kb non-overlapping bins with increasing distance from exons, up to 100kb away.

We find significant positive correlations between our model estimates of both θ_{del} (Spearman's $\rho = 0.47$, $p = 0.00058$) and θ_{ins} (Spearman's $\rho = 0.28$, $p = 0.046$) with distance from exons (figure 3.3). This relationship is corroborated when using π estimates for deletions and insertions (deletions: Spearman's $\rho = 0.79$, $p = 2.2 \times 10^{-16}$, insertions: Spearman's $\rho = 0.84$, $p = 2.2 \times 10^{-16}$, see figure B.6). We separated variants into two γ ranges, 0 to -1 and < -1 and re-analysed this relationship. For the putatively neutral sites we recapture this significant correlation between θ and distance from exons (θ_{del}: Spearman's $\rho = 0.54$, $p = 7.9 \times 10^{-5}$, θ_{ins}: Spearman's $\rho = 0.57$, $p = 2.3 \times 10^{-5}$). However, for the more deleterious category we see no relationship (θ_{del}: Spearman's $\rho = -0.027$, $p = 0.85$, θ_{ins}: Spearman's $\rho = -0.15$, $p = 0.30$) (figure B.7). Additionally, to assess how these correlations held up when using data further from exons we performed correlations on down-sampled datasets by cumulatively removing each bin nearest to exons in turn, progressively reducing our number of bins from 50 to 2. We see that for π we recover significant positive correlations (for both deletions and insertions) for datasets starting up to ∼35kb from exons. For θ we recover this relationship for deletions up to ∼40kb from exons, however for insertions we lack statistical power from the model estimates, probably due to there being relatively fewer insertion polymorphisms (figure B.8).

3.4.4 Recombination rate and INDEL diversity

To obtain additional evidence for linked selection we separated our non-coding INDEL data into 322 2Mb genomic windows, each with a mean recombination rate estimate. As a lack of a regional neutral reference per window precluded the use of our model we instead obtained estimates of π and Tajima's D for each window.

We report positive relationships between π_{ins} and recombination rate (Spearman's $\rho = 0.18$, $p = 0.0010$), and π_{del} and recombination rate (Spearman's $\rho = 0.12$, $p = 0.027$) (figure 3.4a). However, when introducing INDEL divergence as a covariate in a partial correlation analysis (to control for the possible mutagenic effects of recombination), we only maintain the relationship between π_{ins} and recombination rate (partial Spearman's $\rho = 0.15$, $p = 0.0076$) and not π_{del} (patial Spearman's $\rho = 0.077$, $p = 0.17$). Additionally we see a significant enrichment of low frequency variants in low recombining regions, as measured by Tajima's D, for both insertions (Spearman's $\rho = 0.30$, $p = 3.7 \times 10^{-8}$) and deletions (Spearman's $\rho = 0.33$, $p = 1.5 \times 10^{-9}$) (figure 3.4b).

3.5 Discussion

Insertions and deletions often remain unanalysed in sequencing studies, despite constituting a large proportion of genetic variation (Brandstrom and Ellegren, 2007; Montgomery *et al.*, 2013). This is largely a result of the difficulty of working with INDELs compared to SNPs (see Introduction). Yet, when INDELs do get analysed, studies are hampered by the issue of ancestral state misidentification confounding signatures of selection (Kvikstad and Duret, 2014), leaving the selective landscape for INDELs poorly defined. Here we seek to overcome this hurdle using our recently published model (Barton and Zeng, 2018), to estimate the DFE for insertions and deletions in an avian genome. We use high coverage resequencing data from 10 European great tits from Corcoran *et al.* (2017), to quantify the levels of purifying and positive selection for INDELs in coding regions and report evidence of linked selection acting on non-coding INDELs.

3.5.1 Coding sequence INDELs

The majority of INDELs in our dataset are less than 5bp in length. The most common length is 1bp genome-wide, but 3bp within coding regions (figure B.2). This enrichment of in-frame INDELs is even more pronounced in conserved genes (figure B.3). Consistently we report that frame-shifting INDELs have a more severe reduction in diversity and more negative Tajima's D than in-frame INDELs. In non-coding regions we see strong negative correlations between INDEL length and Tajima's D. Taken together these results provide confidence in the genome annotation, show the importance of INDEL length in coding regions with frame-shifting INDELs more deleterious, and provide evidence that longer non-coding INDELs are more deleterious. These results are consistent with previous studies (Barton and Zeng, 2018; Montgomery *et al.*, 2013; Sjödin *et al.*, 2010).

From the application of our model, we see that the majority (96%) of deletions and insertions occurring in CDS regions are strongly deleterious ($\gamma < -100$) (table 3.2, figure 3.2). This proportion corresponds to our previous estimates for INDELs in *Drosophila melanogaster* of between 92% and 97% (Barton and Zeng, 2018). Additionally, our values are similar to those reported for SNPs in a number of organisms, including zerofold degenerate (0-fold) SNPs in the great tit ($\sim$ 80% with $\gamma < -10$) and zebra finch (*Taeniopygia guttata*) ($\sim$ 85% with $\gamma < -10$) (Corcoran *et al.*, 2017), and non-synonymous SNPs in *D. melanogaster* (78% with $\gamma < -100$) and *Mus musculus castaneus* (69% with $\gamma < -100$) (Kousathanas and Keightley, 2013). We estimate the proportion of INDEL substitutions fixed by positive selection, α, at 86% for deletions and 71% for insertions (or 79% and 63% respectively when using non-coding INDELs as neutral reference)(table 3.2). This is comparable to our previous estimates of α for deletions (81%) and insertions (60%) in *D. melanogaster* (Barton and Zeng, 2018), and α estimates for SNPs in *D. melanogaster* of between 74% and 95% (Schneider *et al.*, 2011). However, our estimates are higher than the α estimate for 0-fold SNPs of 48% obtained by Corcoran *et al.* (2017) using the same great tit dataset. This may reflect stronger purifying selection acting on INDELs than SNPs (in line with our Tajima's D and divergence estimates), which provides a stronger opposing force to genetic drift and hence reduces the number of INDEL fixations by drift relative to SNPs. Both our γ estimates for weakly selected sites and α estimates point to deletions being more deleterious than insertions, in line with theoretical expectations that deletions impact more sequence than insertions, and are thus more likely to hit an important motif (Petrov, 2002b; Sjödin *et al.*, 2010), as reported in other studies (Chintalapati *et al.*, 2017; Montgomery *et al.*, 2013; Sjödin *et al.*, 2010).

A number of potential caveats are worth noting however. First, the great tit has likely experienced a recent population expansion (Corcoran *et al.*, 2017; Laine *et al.*, 2016), consistent with our negative Tajima's D values across the genome. Population expansion can lead to an excess of weakly deleterious fixations relative to the amount seen in polymorphism data, which can artificially inflate estimates of the proportion of mutations fixed by positive selection (Eyre-Walker, 2002; Eyre-Walker and Keightley, 2009). Here, we have used the method of Eyre-Walker *et al.* (2006) to control for demography. Existing evidence suggests that this approach is effective in alleviating biases on the estimation of selection intensity on weakly selected variants caused by demography (see Figure 4a in Jackson *et al.*, 2017). Since the best fitting model suggests that the DFE for both insertions and deletions in coding regions is bimodal, with segregating variants subject to weak purifying selection (Table 3.2), our α estimates should be robust.

Second, the formula for estimating α (e.g., eq. 19 in Barton and Zeng, 2018) assumes that the mutation rate is the same between the neutral reference and the focal sites. For this reason, we employed the equal mutation rate model in our analysis of the coding INDELs. However, we note that the model that assumes a gamma DFE and allows the neutral sites and the coding sites to have different mutation rates fits the data better than the equal mutation rate model presented in Table 3.2 [ΔAIC = AIC(best fitting equal mutation rate model) - AIC(best fitting variable mutation rate model) = 4.50]. As demonstrated in Barton and Zeng (2018), this difficulty can be readily alleviated if we know both the point mutation rate and the INDEL mutation rate, which is currently unavailable for the great tit, but can be obtained by direct sequencing of parents and offspring. It should also be noted that both models lead to similar conclusions regarding the DFE. To see this, we calculate $p(|X| \leq x)$ for $x = 1.5$, 5, and 10, where $|X|$ follows a gamma distribution. Using the MLEs (table B.5), for insertions, the proportions are 0.12, 0.18, and 0.23, whereas for deletions, they are 0.052, 0.094, and 0.132. These results are congruent with those shown in Table 3.2 as they indicate that, in coding regions, deletions tend to be under stronger purifying selection, and that only a small fraction

INDEL mutations are sufficiently weakly selected that they contribution to observed polymorphism.

Thirdly as repetitive regions of the genome are notoriously difficult to call variants in and align (Earl *et al.*, 2014), it is possible that our elevated diversity and divergence estimates in ancestral repeats could be the result of an increased number of false positive calls in these regions. To assess the impact of our choice of neutral reference on the DFE we reran our coding analysis using non-coding INDELs as neutral reference. We find that the use of either neutral reference results in a very similar bimodal DFE, with a majority of INDELs being strongly deleterious, and a minority weakly deleterious (Table 3.2). With non-coding INDELs as neutral reference, we observe a slight reduction in the estimated selection pressure on the weakly deleterious site class. This is probably due to the presence of weakly selected variants in the non-coding dataset, as we have previously shown (table B.2, Barton and Zeng, 2018). As the fixation rate is higher when the estimated selection coefficient is smaller, our α estimates are also lower in this case, but are still well above zero. Overall, it seems that our use of ancestral repeats as neutral reference does not unduly impact our results.

3.5.2 Non-coding INDELs and linked selection

The DFE for non-coding INDELs is best described by a gamma distribution. The shape parameter estimates we obtain for both insertions and deletions are small (0.0345 and 0.106 respectively, table 3.2), corresponding to 76% of insertions and 52% of deletions having γ values between 0 and -1, and thus effectively neutral (figure 3.2). The proportion of neutral insertions in non-coding regions (76%) is comparable to the proportion of intronic SNPs with γ estimates between 0 and -1 (70%) in *D. melanogaster* (Eyre-Walker and Keightley, 2009). However, the proportion of deletions falling into this selective range is markedly lower at 52%, more in line with SNPs in untranslated regions in birds, where in the great tit $\sim$ 50%, and in the zebra finch $\sim$ 40% of variants fall within the 0 to -1 γ range (Corcoran *et al.*, 2017). This mirrors and reinforces the trend seen in coding regions supporting the more deleterious nature of deletions. It also suggests that overall a substantial proportion of INDELs (24% of insertions and 48% of deletions) in non-coding regions are experiencing purifying selection.

To understand how non-coding INDEL diversity changes around coding regions, we investigated how θ varies with distance from exons. Our analysis shows that non-coding θ estimates adjacent to exons are lower than the genome-wide non-coding estimates. As distance from exons increases, both θ_{ins} and θ_{del} increase significantly returning to the genome-wide level by 100kb from exons (figure 3.3). As the scaled mutation rate ($\theta = 4N_e\mu$) is the product of the per site mutation rate (μ) and the effective population size (N_e) changes in θ can be the result of changes in either parameter. However, as we do not expect there to be a systematic variation in μ between our distance bins, changes in θ should be driven by corresponding changes in N_e. This relationship between distance and θ could be explained through increasing proximity to functional sequence, and therefore increased linkage to sites either under purifying or positive selection, resulting in reduced N_e close to exons (see Cutter and Payseur (2013) for review). Alternatively, it could be driven by a higher density of regulatory elements under selective constraint in non-coding sequence near exons, making INDELs closer to exons more deleterious, and thus reducing diversity in these regions. However, two lines of evidence presented here support the former explanation. Firstly, we can recapture the relationship between INDEL diversity and distance from exons when re-analysing our dataset after removing data up to as much as the nearest 30kb to exons for π_{ins}, π_{del} and θ_{del} (although for θ_{ins} we lack statistical power). This demonstrates that the correlation is not solely driven by regions directly neighbouring exons, as might be expected if driven by purifying selection on regulatory elements, but extends over larger distances, more indicative of linked selection (figure B.8). Secondly, when we analyse nearly neutral variants ($-1 \leq \gamma \leq 0$) and deleterious variants ($\gamma < -1$) separately we see that the relationship between distance from exons and θ is driven by a significant increase in nearly neutral variants as distance from exons increases. We see no increase in deleterious variants close to exons as would be expected if regulatory elements were disrupted (figure B.7). Additionally, this suggests that while a proportion of INDELs in non-coding regions seem to be experiencing negative selection, in agreement with our reported genome-wide non-coding DFE, these variants are not driving the reduction of diversity in proximity to exons.

The possibility of linked selection reducing diversity is further supported by the significant positive correlations we see between local recombination rate and π_{ins}, π_{del} and Tajima's D (figure 3.4). Linked selection can be expected to generate such a pattern, with linkage decreasing as recombination rates increase, which should drive higher π in high recombining regions (Corcoran *et al.*, 2017) and a greater enrichment of low frequency variants in low recombining regions. However, the mutagenic effect of recombination can also be expected to generate a relationship between π and recombination (Arbeithuber *et al.*, 2015). To disentangle these two forces, we conducted partial correlation analyses using INDEL divergence as a covariate. The partial correlation coefficient between π_{ins} and recombination is 0.15, which is significant and close to the value of 0.18 obtained without using divergence as a covariate. In contrast, the partial correlation coefficient between π_{del} and recombination rate is 0.077, which is non-significant and more different from the value of 0.12 obtained without partial correlation. This suggests that the mutagenic effect of recombination has probably played a role in driving increased INDEL mutation rates in high recombining regions, and that this effect is likely stronger for deletions than insertions. This is in line with results previously reported in zebra finch (Nam and Ellegren, 2012). Yet, the greater enrichment in low frequency variants in low recombining regions is not an expected outcome of reduced mutation rates. Thus, it seems likely that the true picture is a combination of both linked selection and mutation variation shaping patterns of INDEL variability in regions of varying recombination.

3.5.3 Conclusion

In summary, we see that genome-wide INDELs appear to be having detrimental effects, with most coding INDELs strongly deleterious, and a sizeable minority of non-coding INDELs showing signatures of purifying selection. We also show that non-coding INDEL diversity is constrained through linkage to selected sites near exons and in low recombining regions, though some of this can be attributed to the mutagenic effect of recombination. However, we cannot separate how much of this trend is driven by positive selection and how much is due to purifying selection, which would be an interesting avenue for future INDEL studies.

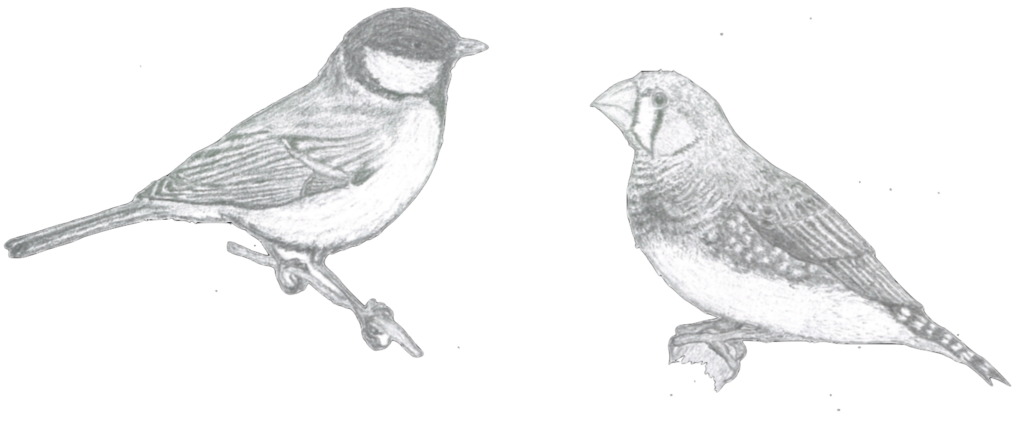
3.6 Acknowledgements

We thank Pádraic Corcoran for advice on the variant calling pipeline and assistance with the SNP data and Alison Wright for suggestions on the investigation of selection at linked sites. This work was supported by a PhD studentship funded by the Department of Animal and Plant Sciences, University of Sheffield, to H.J.B. Support was also provided by the Natural Environment Research Council via a research grant awarded to K.Z. (NE/L005328/1).

Chapter 4

The impact of biased gene conversion and insertion and deletion variation on GC content in two passerines

Authors: **Henry J. Barton** and **Kai Zeng**



4.1 Abstract

Understanding the determinants of genomic base composition is fundamental to understanding genome evolution. GC biased gene conversion (gBGC) is a key driving force behind genomic GC content through the preferential incorporation of GC alleles over AT alleles during recombination, driving them to fixation. To date, the majority of work on gBGC has focussed on its role in coding regions, largely to address how it confounds estimates of selection. More generally, the evolution of base composition has predominantly been viewed from the perspective of point mutations. To address these biases, we investigate how the strength of gBGC (B) varies within the non-coding genome of two wild passerines. We also characterise the impact of both polymorphic and fixed small INDELs on genomic base composition. Using a dataset of 20 previously published high coverage genomes (10 great tits and 10 zebra finches) we estimate recombination rate and B in 1Mb homologous windows in each species. We demonstrate remarkable conservation of both B and recombination between species. Additionally, we show the mean strength of gBGC in the zebra finch is more than double that in the great tit, consistent with its twofold greater effective population size. We estimate equilibrium GC content from both divergence and polymorphism data which indicates that equilibrium GC content has increased as a result of recent population expansions in both species. Finally we show that neither polymorphic or fixed INDELs are GC conservative in nature and have the ability to shape genomic base composition.

4.2 Introduction

A large proportion of many organisms' genomes are non-coding; 99% in humans, 91% in *Drosophila melanogaster*, 73% in *Caenorhabditis elegans* and 71% in *Arabidopsis thaliana* (Rajic *et al.*, 2005). The non-coding genome offers the opportunity to study evolutionary processes away from the interference of the direct impacts of natural selection, and can allow us to study forms of variation, such as insertions and deletions, that segregate at too low a frequency in coding regions to address many questions statistically. One such process is the evolution of genomic base composition. The evolution of base content and its variation within genomes has been the focus of intrigue for many years, such as the question of mammalian isochore evolution (Eyre-Walker and Hurst, 2001). Predominantly, research into the evolution of genomic GC content has focussed on the balance between the strong to weak substitution rate (S→W), in part underpinned by CpG hypermutabiliy (Hodgkinson and Eyre-Walker, 2011; Hwang and Green, 2004; Ségurel *et al.*, 2014), and the weak to strong substitution rate (W→S), which is heavily influenced by the role of GC biased gene conversion (gBGC) which has been shown to be a major determinant of GC content evolution in a broad range of organisms (Bolívar *et al.*, 2016, 2018, 2019; Corcoran *et al.*, 2017; Glémin *et al.*, 2015; Gossmann *et al.*, 2018; Jackson *et al.*, 2017; Muyle *et al.*, 2011; Ratnakumar *et al.*, 2010; Wallberg *et al.*, 2015).

The process of gBGC is the preferential incorporation of GC alleles over AT alleles during the resolution of heteroduplex DNA resulting from the repair of double stranded breaks during recombination. This elevates the number of gametes containing GC alleles, as observed in humans (Williams *et al.*, 2015) and birds (Smeds *et al.*, 2016). As such, gBGC acts to increase the frequency of G and C alleles over A and T alleles, in a manner that mirrors positive selection (Duret and Galtier, 2009; Galtier and Duret, 2007). As a result, gBGC is an inconvenient complication when looking for signatures of selection in genomes. For example over 20% of identified positively selected genes in the human lineage are possibly instead just the focus of elevated gBGC (Ratnakumar *et al.*, 2010). Furthermore, a growing body of literature has demonstrated that gBGC confounds our ability to estimate parameters such as the rate of adaptation ($\omega = dN/dS$) (Bolívar *et al.*, 2018, 2019; Corcoran *et al.*, 2017; Gossmann *et al.*, 2018; Ratnakumar *et al.*, 2010; Rousselle *et al.*, 2019) and the proportion of substitutions fixed by positive selection (α) (Bolívar *et al.*, 2018; Corcoran *et al.*, 2017; Rousselle *et al.*, 2019). Equally this can be framed as studying gBGC in coding regions is inconvenienced by the action of natural selection also acting on those regions, forcing studies to use putatively neutral sites like third codon positions (Rousselle *et al.*, 2019; Weber *et al.*, 2014) and 4-fold degenerative sites (Bolívar *et al.*, 2016; Corcoran *et al.*, 2017; Gossmann *et al.*, 2018) reducing the amount of data available as well as potentially being confounded by codon usage bias (Galtier *et al.*, 2018; Jackson *et al.*, 2017).

As gBGC is a recombination mediated process, it has the potential for as much variation in strength as recombination rate, at different genomic scales and between species, as supported by a large body of literature demonstrating correlations between recombination rate and GC content (Bolívar *et al.*, 2016; Glémin *et al.*, 2015; Rousselle *et al.*, 2019; Wallberg *et al.*, 2015; Weber *et al.*, 2014), recombination rate and equilibrium GC content (GC*) (Duret and Arndt, 2008; Muyle *et al.*, 2011; Singhal *et al.*, 2015), and recombination rate and the population scaled strength of gBGC, $B = 4N_e b$, where N_e is the effective population size and b is the raw strength of conversion bias (Glémin *et al.*, 2015; Wallberg *et al.*, 2015). With recombination processes varying greatly between organisms (Stapley *et al.*, 2017) this can be expected to give a similarly broad range of impacts and strength of gBGC. For example, in mammals the recombination landscape is largely determined by the location of recombination hotspots, determined by the PRDM9 gene (Baudat *et al.*, 2010; Parvanov *et al.*, 2010). This results in areas of greatly elevated recombination rate and thus strength of gene conversion relative to background levels, for example in humans mean B is estimated at ~ 0.4 (Glémin *et al.*, 2015), while inside recombination hotspots it reaches as high as ~ 18 (Glémin *et al.*, 2015). In birds, the combination of a karyotype consisting of a few long macro-chromosomes and many smaller micro-chromosomes (Hansson *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014; Zhang *et al.*, 2014) and obligate crossing over causes large chromosomal differences in recombination rate (Backström *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014). Additionally, it has been suggested that birds' lack of PRDM9, has resulted in stable recombination hotspots and conserved recombination characteristics between species (Singhal *et al.*, 2015). Together this is suggested to allow strong gBGC to act on the same region of the genome over a longer time period than in mammals (Rousselle *et al.*, 2019; Singhal *et al.*, 2015), driving GC content increases, with studies reporting that GC content is below GC* content in most avian lineages (Bolívar *et al.*, 2016; Mugal *et al.*, 2013; Rousselle *et al.*, 2019; Weber *et al.*, 2014). Furthermore, some organisms, such as the honey bee *Apis mellifera*, lack pronounced recombination hotspots, yet have very high genome-wide recombination rate with 5 crossovers per arm and correspondingly elevated mean B estimates of ~ 5 (Wallberg *et al.*, 2015). Overall, gBGC is seemingly an ubiquitous force with mean B estimates varying from 0.4 to 5 across the tree of life (Long *et al.*, 2018).

As B is defined as $4N_eb$, not only is its strength modulated by recombination rate increasing b (the strength of conversion) as outlined above but also by the effective population size (N_e). As such species with larger N_e should have larger B and a reduced impact of genetic drift. This has been reported in a few studies, with correlations between N_e and GC content at 3rd codon positions (GC3) in birds, largely driven by increased GC in smaller bodied, larger N_e species, as well as correlations between N_e and GC* (Weber *et al.*, 2014). More recently B at fourfold degenerate sites (4-fold sites) has been shown to correlate with N_e in great apes (preprint: Borges *et al.*, 2018). However, analysis of B more broadly across animal taxa, failed to yield a relationship between B and N_e (Galtier *et al.*, 2018). Furthermore, to date the role of N_e is a less well empirically studied aspect of gBGC and little work has looked at fine scale variation in the strength of gBGC between species of differing N_e.

Another intriguing, but so far largely un-addressed, potential contributor to GC content evolution is small insertions and deletions (small INDELs, here $\leq 50bp$). As the fixation of small insertions results in the addition of bases and the fixation of deletions the loss of nucleotides, differences between deletion and insertion base content, and/or differences in fixation rate between insertions and deletions, have the potential to influence base composition. Indeed an insertion fixation bias has been reported in a number of studies (Chintalapati *et al.*, 2017; Leushkin and Bazykin, 2013; Presgraves, 2006; Sjödin *et al.*, 2010), although the root cause of this relationship is a source of contention, with a number of explanations put forward, including selection on insertions to prevent intron length decreasing below an optimum (Ometto *et al.*, 2005; Parsch, 2003; Presgraves, 2006), insertion biased gene conversion (Leushkin and Bazykin, 2013) and ancestral state misidentification (Kvikstad and Duret, 2014). Secondly, chicken INDELs have been seen to be enriched in A containing motifs (Brandstrom and Ellegren, 2007), and there is some limited evidence that the ratio of deletions to insertions is greater in GC rich introns in humans (Wang and Yu, 2011). The interaction of relative deletion and insertion rates and variable INDEL base content has the potential to influence genomic base composition, and warrants further investigation.

The avian system has been the model of choice for many studies addressing GC evolution and biased gene conversion (Bolívar *et al.*, 2016, 2018, 2019; Corcoran *et al.*, 2017; Gossmann *et al.*, 2018; Rousselle *et al.*, 2019; Weber *et al.*, 2014) as well as addressing questions relating to small INDELs (Barton and Zeng, 2019; Boschiero *et al.*, 2015; Brandstrom and Ellegren, 2007; Johnson, 2003; Nam and Ellegren, 2012; Paśko *et al.*, 2011; Sundström *et al.*, 2003; Yan *et al.*, 2014). The suitability of avian genomes for addressing these topics stems from their variable intra genomic recombination landscapes (Backström *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014) and conserved recombination hotspots (Singhal *et al.*, 2015) providing a natural experiment for addressing the role of recombination and N_e in gBGC and GC content evolution. In addition birds' conserved karyotype and synteny (Hansson *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014; Zhang *et al.*, 2014) allow for straightforward whole genome alignments to create orthologous window datasets and infer ancestral states and polarise variants such as small INDELs.

To date most work has focussed on exploring the impact of gBGC and its interaction with the above described genomic traits within genes and coding regions, largely with a view of addressing how it confounds signatures of selection (Bolívar *et al.*, 2019; Corcoran *et al.*, 2017; Gossmann *et al.*, 2018; Ratnakumar *et al.*, 2010; Rousselle *et al.*, 2019). As such, there has been little attention paid to the action of gene conversion in the non-coding genome, although there are some notable exceptions (Glémin *et al.*, 2015; Jackson *et al.*, 2017; Muyle *et al.*, 2011; Wallberg *et al.*, 2015), and little work investigating fine scale variation in gBGC across the genome between species. Furthermore the role of small INDELs in the evolution of GC content remains un-addressed. Here we investigate variation in the strength of gBGC within the non-coding genome of two passerine species, the great tit (*Parus major*) and the zebra finch (*Taeniopygia guttata*), using previously published whole genome resequencing data (Corcoran *et al.*, 2017; Singhal *et al.*, 2015). We seek to address how conserved the gBGC landscape is between these species and how gBGC has influenced the evolution of base composition in these lineages. Additionally, we attempt to characterise how small INDELs have influenced base composition dynamics in these birds.

4.3 Materials and Methods

4.3.1 The dataset

The dataset consisted of 10 European great tits from across the sampling locations in Laine *et al.* (2016), sequenced to a mean coverage of 44X and 10 Australian zebra finches sequenced to a mean coverage of 22X, a subset of ten individuals from the Fowlers Gap population in Australia from the dataset published in Singhal *et al.* (2015). The dataset is as described in Corcoran *et al.* (2017). For both species we obtained VCF files for SNPs and monomorphic sites from Corcoran *et al.* (2017). A VCF file for small INDELs ($\leq 50bp$) for the great tit dataset was obtained from Barton and Zeng (2019). For the zebra finch we downloaded INDEL realigned and base quality score recalibrated BAM files for each of the ten zebra finch, prepared as described by Singhal *et al.* (2015), from http://www.ebi.ac.uk/ena/data/view/PRJEB10586 (last accessed 05/03/19) and called insertions and deletions following the pipeline in Barton and Zeng (2019). Additionally, a three species whole genome alignment between zebra finch, great tit and flycatcher (*Ficedula albicollis*) was obtained from Barton and Zeng (2019), and a three species alignment between chicken (*Gallus gallus*), zebra finch and great tit from Corcoran *et al.* (2017).

4.3.2 Annotation and filtering

We assigned the ancestral states for the SNPs and INDELs using the whole genome alignment and parsimony based approach, where for each species either the reference allele or the alternate allele had to supported by both out-groups to be assigned as ancestral.

We downloaded the great tit genome annotation (version 1.03) from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/522/545/GCF_001522545.1_Parus_major1.0.3/GCF_001522545.1_Parus_major1.0.3_genomic.gff.gz (last accessed 05/03/19) and the zebra finch annotation from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/151/805/GCF_000151805.1_Taeniopygia_guttata-3.2.4 (last accessed on 05/03/19).

We used the annotations to remove variants falling within exons. Additionally coordinates for ultra-conserved non-coding elements (UCNEs) in the zebra finch genome (taeGut1) were obtained from ftp://ccg.vital-it.ch/UCNEbase/custom_tracks_UCSC/UCNEs_taeGut1.bed (last accessed 05/03/19). We identified the corresponding positions in the great tit in the whole genome alignment, before removing any variants falling within UCNEs. Additionally we restricted our analysis to the autosomes, removing the Z chromosome. This left non-coding datasets of putatively neutral variants, numbering 9,800,315 SNPs plus 1,096,350 INDELs for great tit, and 29,973,954 SNPs plus 3,587,939 INDELs for zebra finch.

From our non-coding SNP dataset we generated an additional subset, with CpG sites excluded, where a CpG site was defined as any site where at least one of the alleles of the site was in a $5' \rightarrow 3'$ CpG dinucleotide or in a $3' \rightarrow 5'$ GpC dinucleotide.

4.3.3 Orthologous window preparation

The zebra finch genome was divided into 1Mb non-overlapping windows and we used the three species whole genome alignment (zebra finch, great tit, flycatcher) to identify the aligned sequence and coordinates in the great tit genome and extracted variants and numbers of callable sites from our VCF files. For each window in each species we calculated the GC content using the respective reference genomes. GC content was calculated for all sites in the window, and for non-CpG sites. Secondly, we calculated recombination rate for each window, using the available linkage map data for each species (Stapley *et al.*, 2008; van Oers *et al.*, 2014, for zebra finch and great tit respectively) and the pipeline outlined in Corcoran *et al.* (2017).

4.3.4 Estimating the strength of gene conversion

We extracted the number of callable site for weak bases (A and T nucleotides) and strong bases (G and C nucleotides) along with the site frequency spectra for weak to strong mutations (WS), strong to weak mutations (SW) and weak to weak and strong to strong mutations ($WWSS$) in all windows and datasets. We then applied the $M1*$

model of Glémin *et al.* (2015), implemented in the anavar package (Barton and Zeng, 2018). Briefly, the model estimates the population scaled mutation rate ($\theta = 4N_e\mu$), the population scaled selection coefficient ($\gamma = 4N_es$) and estimates and controls for polarisation error for both SW and WS mutations using $WWSS$ sites as a neutral reference unaffected by gBGC. The selection coefficient for WS mutations (γ_{WS}) can also be thought of as the population scaled strength of biased gene conversion, B, where $B = 4N_eb$.

We performed partial correlations of B against GC content, and recombination rate, across windows, with nucleotide diversity (π) (Tajima, 1983) as a confounding factor to control for the contribution of effective population size (N_e) to the correlation. The partial correlations were performed using the 'ppcor' package (Kim, 2015) in R (R Core Team, 2015).

4.3.5 Equilibrium GC content

We estimated the ancestral GC content per window for the lineage leading to great tits and zebra finches using the whole genome alignment (containing chicken, zebra finch and great tit) and the GTR-NH$_b$ model in baseml within PAML (Yang, 2007). The model allows for non-stationary base content and for independent substitution rates on each branch. From the model we obtained the posterior probabilities of the ancestral states and weighted each ancestral nucleotide by this probability (as in Matsumoto *et al.*, 2015) to reconstruct ancestral GC content with uncertainty incorporated. We then estimated the rate of WS substitutions

$$r_{WS} = \frac{n_{WS}}{n_W} \tag{4.1}$$

the rate of SW substitutions

$$r_{WS} = \frac{n_{SW}}{n_S} \tag{4.2}$$

and the equilibrium GC content

$$GC{*}_{div} = \frac{r_{WS}}{r_{WS} + r_{SW}} \tag{4.3}$$

The GTR-NH$_b$ model was a better fit then the GTR model, which assumes base composition is at equilibrium, for all but five windows as judged by likelihood ratio tests (data not shown). Additionally, the model estimates of GC$*_{div}$ correlated strongly with those derived from parsimony estimates of the substitution rates for both great tit (Pearson's $r = 0.94$, $p < 2.2 \times 10^{-16}$) and zebra finch (Pearson's $r = 0.96$, $p < 2.2 \times 10^{-16}$), although the mean GC$*_{div}$ was lower for the model estimates than the parsimony estimates in both species (0.39 versus 0.43 respectively for great tit and 0.38 versus 0.42 respectively for zebra finch). Additionally we calculated the distance from equilibrium GC content for each window, where distance from equilibrium was defined as: $GC *_{div} - GC_{current}$.

To obtain a more recent view of the base composition evolution and gBGC we also calculated GC$*_{pol}$ from our application of the Glémin *et al.* (2015) model to our polymorphism dataset. This was achieved using the selection coefficients and mutation rates estimated per window to estimate the fixation rates for $W \rightarrow S$ and $S \rightarrow W$ mutations and to substitute these into equation 4.3

$$r_{ij} = \theta_{ij} \frac{\gamma_{ij}}{1 - e^{-\gamma_{ij}}} \tag{4.4}$$

where r_{ij} is the fixation rate of mutations from i to j.

4.3.6 INDEL base content

To assess if there were any GC or AT trends in insertion and deletion content, we estimated the scaled mutation rate ($\theta = 4N_e\mu$) for small INDELs ($\leq 50bp$) consisting of only GC nucleotides, those containing only AT nucleotides, and the remainder of variants which contained a mix of AT and GC nucleotides. For each group of INDELs we obtained the site frequency spectra for non-coding, UCNE filtered, insertions and deletions. For each of the three categories we fit the 'neutralSNP_vs_selectedINDEL' model with 1 site class ($C = 1$) in the anavar package (Barton and Zeng, 2018), using WWSS SNPs as neutral reference (as in the biased gene conversion analysis described earlier). The model provides estimates of the mutation rate after correcting for ancestral state misidentification when separating INDELs into insertions and deletions, and controls for demography using the set of neutral variants. We also calculated diversity (π) for this dataset (Tajima, 1983).

To see how much small INDELs have contributed to the evolution of base composition since the common ancestor of the great tit and zebra finch we estimated the change in GC content due to small insertion fixations (ΔGC_{ins})

$$\Delta GC_{ins} = \sum \frac{GC_{anc} + GC_{ins}}{l + l_{ins}} - \frac{GC_{anc}}{l} \tag{4.5}$$

where GC_{anc} is the ancestral GC content, GC_{ins} is the insertion GC content, l is the ancestral sequence length and l_{ins} is the insertion length. Similarly, for deletion fixations we estimated ΔGC_{del}

$$\Delta GC_{del} = \sum \frac{GC_{anc} - GC_{del}}{l - l_{del}} - \frac{GC_{anc}}{l} \tag{4.6}$$

where GC_{del} is the deletion GC content and l_{del} is the deletion length. We calculated ΔGC_{ins} and ΔGC_{del} for each window from our whole genome alignment dataset. We removed windows with fewer than $200bp$ of INDELs.

4.3.7 Data availability

All scripts and command lines used in the analysis pipeline can be found at: `https://github.com/henryjuho/biased_gene_conversion`.

Table 4.1: Summary of the window dataset, showing means and the 2.5 and 97.5 percentiles in brackets. Recombination rates are log10 transformed.

Measure	great tit	zebra finch
windows	898	904
callable sites	523858 (21580, 726488)	498785 (79743, 711346)
n_{SNP}	5895 (239, 9766)	21321 (989, 37847)
n_{INS}	257 (10, 484)	869 (46, 1447)
n_{DEL}	331 (3, 661)	1602 (71, 2874)
GC content	0.41 (0.34, 0.51)	0.41 (0.35, 0.51)
Recombination rate (cM/Mb)	0.48 (0, 0.97)	0.41 (0, 0.96)

Table 4.2: Results of partial Spearman's correlations of the strength of gene conversion (B) with GC content and recombination rate, with π as a covariate, using both the main dataset and the dataset after removing CpG sites.

x	y	dataset	species	Spearman's ρ	p value
Recombination rate (cM/Mb)	B	full	great tit	0.43	9.09×10^{-40}
Recombination rate (cM/Mb)	B	full	zebra finch	0.56	6.02×10^{-74}
Recombination rate (cM/Mb)	B	CpG filtered	great tit	0.50	1.15×10^{-52}
Recombination rate (cM/Mb)	B	CpG filtered	zebra finch	0.55	1.49×10^{-70}
GC content	B	full	great tit	0.50	9.61×10^{-55}
GC content	B	full	zebra finch	0.75	9.18×10^{-157}
GC content	B	CpG filtered	great tit	0.54	2.2×10^{-62}
GC content	B	CpG filtered	zebra finch	0.75	2.46×10^{-158}

4.4 Results

4.4.1 Summary of the window dataset

Application of our pipeline resulted in 904 1Mb windows in zebra finch genome and 898 orthologous windows in the great tit genome (table 4.1). The lower number of great tit windows is due to gaps in the whole genome alignment. We see similar numbers of callable sites in both species, roughly 500,000 bp per 1 Mb window, this drop is a result of our maximum parsimony approach to assigning ancestral states, which is dependant on coverage of all species in our whole genome alignment and no ambiguity between out-groups. When considering variants per window we see that the mean number of variants is higher in zebra finch for all mutation classes (SNPs, insertions and deletions), consistent with a larger effective population size in zebra finch (Corcoran *et al.*, 2017). We see very similar mean GC content and mean recombination rates in both species, with strong correlations between the two species' GC content (Pearson's $r = 0.83$, $p = 1.6 \times 10^{-230}$, figure C.1a) and recombination rate (Spearman's $\rho = 0.72$, $p = 2.6 \times 10^{-140}$,

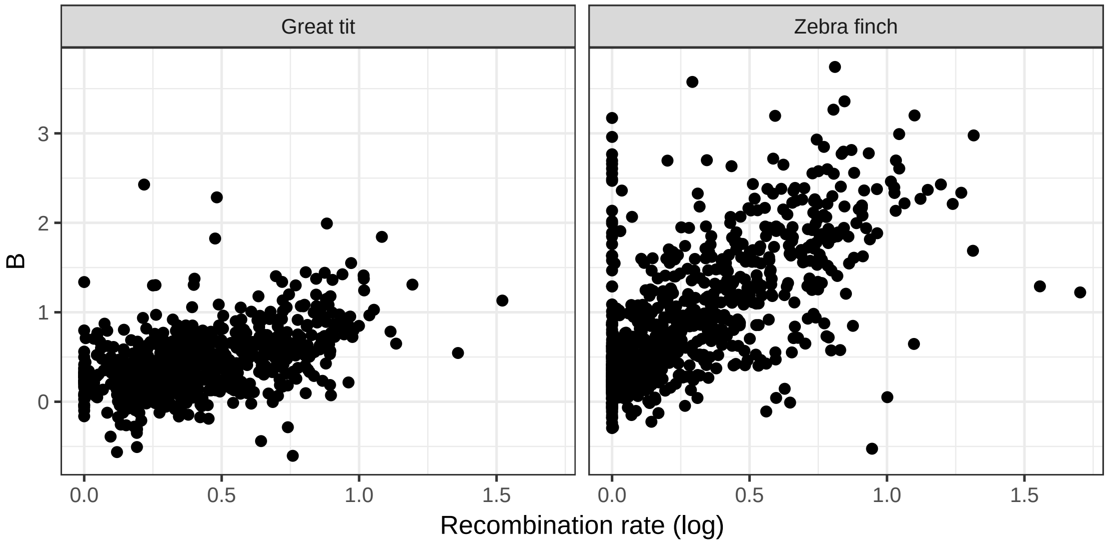


FIGURE 4.1: The relationship between mean window recombination rate and the strength of gene conversion (B) in the great tit and zebra finch. Spearman's correlation results can be seen in the top two rows of table 4.2.

figure C.1b) across the dataset, as well as positive correlations between GC content and recombination within each species (great tit: Spearman's $\rho = 0.57$, $p = 3.8 \times 10^{-79}$, zebra finch: Spearman's $\rho = 0.53$, $p = 4.2 \times 10^{-67}$, figure C.2).

4.4.2 Strength of gene conversion correlates with recombination

The strength of GC-biased gene conversion (B) positively correlates with recombination rate in both the great tit and the zebra finch (table 4.2, figure 4.1). This relationship is stronger when using mean GC content as a proxy for recombination rate in both species (table 4.2, figure C.3) and all correlations are maintained when performed on a dataset filtered for CpG sites (table 4.2).

4.4.3 *B* conserved between the species

Comparison of the model estimates of B between zebra finch and great tit show a significantly larger mean B value in zebra finch ($\bar{B} = 0.90$) than great tit ($\bar{B} = 0.40$) (Wilcoxon rank sum, $W = 491903$, $p = 2.5 \times 10^{-49}$; figure 4.2a). However, when we standardise our B estimates by π as a measure of N_e the distributions are similar

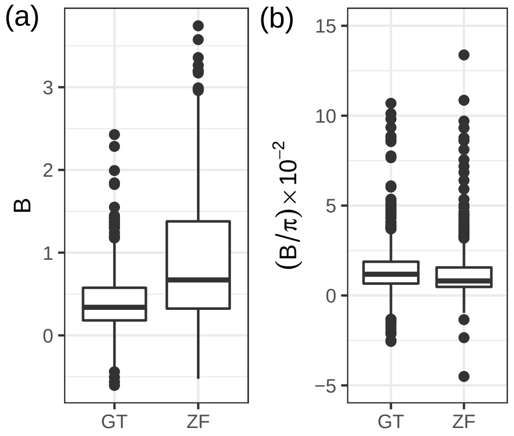


FIGURE 4.2: Comparison of the distribution of B values (strength of biased gene conversion) (a) and B standardised by π as a proxy of the effective population size N_e (b) between the great tit (GT) and zebra finch (ZF). The y axis for b has been cropped for clarity.

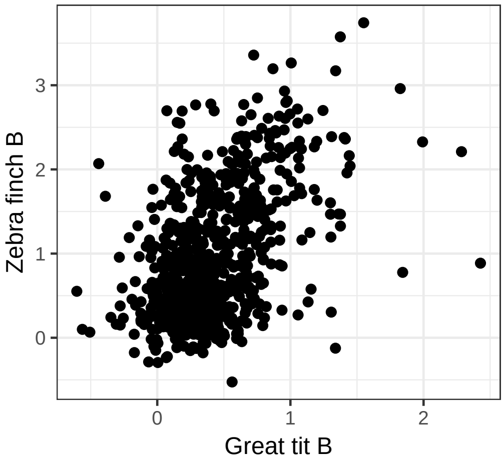


FIGURE 4.3: The relationship between the strength of biased gene conversion (B) in the zebra finch and the great tit.

between the two species (figure 4.2b), although the species' means remain significantly different (Wilcoxon rank sum, $W = 305880$, $p = 6.1 \times 10^{-10}$). We also see a positive correlation between the ratio of the species' nucleotide diversity (π_{ZF}/π_{GT}) and the ratio of the species' B (B_{ZF}/B_{GT}) (Spearman's $\rho = 0.44$, $p < 2.2 \times 10^{-16}$), supporting the idea that the N_e drives the between species differences in B. Furthermore, we see a strong correlation between B in the great tit and B in the zebra finch (Spearman's $\rho = 0.45$, $p = 9.84 \times 10^{-42}$, figure 4.3) as well as between B/π in great tit and B/π in zebra finch (Spearman's $\rho = 0.47$, $p = 1.91 \times 10^{-46}$), in keeping with the conserved recombination rate and GC content between species reported above.

4.4.4 Equilibrium GC content

To assess the longer term GC dynamics of both the great tit and zebra finch genomes we calculated the equilibrium GC content (GC$*_{div}$) for each lineage using the substitution rates estimated in PAML (see methods). This gave a mean GC$*_{div}$ of 0.39 for great tit and 0.38 for zebra finch, both of which are significantly below the mean GC contents in our alignment datasets of 0.40 for both great tit (Wilcoxon rank sum, $W = 282790$, $p =$

Table 4.3: Correlations between the distance from equilibrium GC content (GC$*_{div}$ - GC) and other genomic variables across the window dataset.

variable	species	correlation	p value	method
Recombination rate	great tit	0.21	1.83×10^{-14}	Spearman's
Recombination rate	zebra finch	0.59	9.95×10^{-137}	Spearman's
B	great tit	0.28	1.18×10^{-23}	Spearman's
B	zebra finch	0.61	9.96×10^{-151}	Spearman's

1.1×10^{-8}) and zebra finch (Wilcoxon rank sum, $W = 241190$, $p < 2.2 \times 10^{-16}$) (figure C.7). Note the alignment dataset is a subset of the main dataset (as coverage is required across all species in the chicken/zebra finch/great tit alignment) and yields slightly lower mean GC. B positively correlates with GC$*_{div}$ in both great tit (Spearman's $\rho = 0.50$, $p < 2.2 \times 10^{-16}$) and zebra finch (Spearman's $\rho = 0.87$, $p < 2.2 \times 10^{-16}$), and a similar relationship is seen for recombination (Spearman's $\rho = 0.55$, $p = 6.02 \times 10^{-62}$ for great tit and Spearman's $\rho = 0.66$, $p = 3.85 \times 10^{-98}$ for zebra finch).

To look at base composition evolution in a more recent time scale we also calculated equilibrium GC content using our θ and B estimates derived from the polymorphism data (see methods), henceforth GC$*_{pol}$. This approach yielded markedly higher equilibrium GC content estimates than the substitution rate based approach, for both great tit (Wilcoxon rank sum, $W = 518421$, $p = 1.48 \times 10^{-225}$, $G\bar{C}*_{pol} = 0.63$) and zebra finch (Wilcoxon rank sum, $W = 575196$, $p = 1.24 \times 10^{-245}$, $G\bar{C}*_{pol} = 0.72$).

4.4.5 Distance from equilibrium GC

To further quantify the impact of biased gene conversion on GC content, we calculated the distance from equilibrium (GC$*_{div}$) for each window, in each species. This results in a mean distance from equilibrium of -0.0034 in great tit and -0.015 in zebra finch (also see figure C.7). Additionally, both B and recombination rate positively correlate with the distance from equilibrium in both species (table 4.3, figure 4.4, figure C.8). This suggests that regions that substitution based analysis predicts to increase in GC content (i.e., $GC *_{div} - GC > 0$) are also under stronger recent selection, with larger B values, supporting our use of the longer term GC$*_{div}$ over the shorter term GC$*_{pol}$ in this analysis. As the distance from equilibrium ranges from values below equilibrium to values

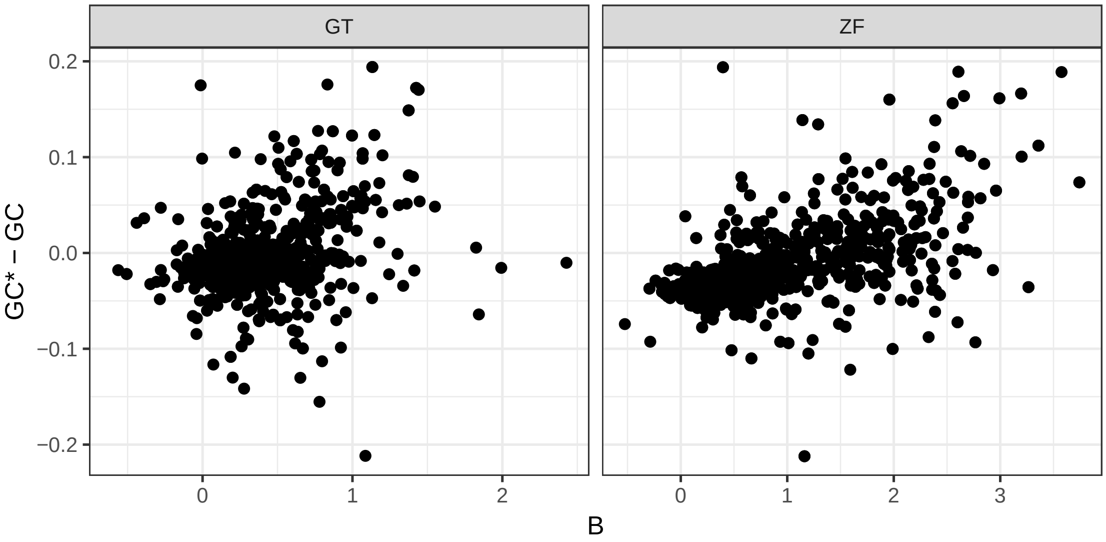


Figure 4.4: The relationship between distance from equilibrium GC (GC$*_{div}$ - GC) and the strength of biased gene conversion (B) in both the great tit (GT) and the zebra finch (ZF).

above equilibrium we binned the estimates into 3 categories, below equilibrium, above equilibrium and at equilibrium (current GC content within 2.5% of GC$*_{div}$) to assess how much these relationships were driven by the above equilibrium and below equilibrium windows. This binning suggests that the correlations between B and distance from equilibrium (table 4.3) is largely driven by elevated B estimates in windows with below equilibrium GC content in both species (figure C.9a). Similarly the correlations between recombination rate and distance from equilibrium GC (table 4.3) is largely driven by higher recombination rates in windows with below equilibrium GC (figure C.9b).

4.4.6 Impact of small INDEL mutations on GC

To understand how INDEL mutation rates might be influencing base content we estimated the population scaled mutation rate ($\theta = 4N_e$) for INDELs containing only GC bases, INDELs containing only AT bases and INDELs containing both AT and GC bases (mixed INDELs). Mean GC content of mixed INDELs (0.43 in great tit, 0.45 in zebra finch, figure C.4) is slightly higher than the mean non-coding GC content (table 4.1). The nature of our INDEL binning means that AT and GC INDELs share similar length distributions (predominantly comprising of 1bp INDELs), whereas mixed INDELs differ

Table 4.4: Correlations between INDEL ΔGC and ancestral GC content across the window dataset.

species	INDEL type	correlation	p value	method
great tit	INS	0.74	4.23×10^{-127}	pearson
zebra finch	INS	0.67	4.64×10^{-97}	pearson
great tit	DEL	-0.81	2.27×10^{-171}	pearson
zebra finch	DEL	-0.89	4.04×10^{-251}	pearson
great tit	INDEL	0.18	9.08×10^{-7}	pearson
zebra finch	INDEL	-0.65	6.76×10^{-89}	pearson

as they must be at least 2bp long to contain a mix of GC and AT bases, and constitute the majority of the overall distribution of INDEL lengths above 1bp (figure C.5). We see that for both GC and mixed INDELs θ is greater for deletions than insertions in both species (figure 4.5). However, AT INDELs show a strikingly contrasting pattern with θ_{ins} exceeding θ_{del}, where $\theta_{del}/\theta_{ins} = 0.84$ in the great tit, and θ_{ins} is a very similar magnitude to θ_{del} in the zebra finch with $\theta_{del}/\theta_{ins} = 1.1$. For comparison the ratios of $\theta_{del}/\theta_{ins}$ for GC and mixed INDELs are 1.8 and 2.2 respectively in the great tit and both 3.2 in the zebra finch. These relationships are corroborated when using π estimates (figure C.6). Our estimates of the population scaled selection coefficient ($\gamma = 4N_es$) show that in both species GC containing and mixed INDELs are characterised by deletions with negative γ estimates and insertions with positive γ estimates (table C.1). Conversely, for AT containing variants deletions have more positive γ values than insertions, however these estimates are both negative in the great tit and both positive in the zebra finch (table C.1).

4.4.7 GC content change from INDEL fixation

To quantify how small INDELs have influenced GC content since the common ancestor of the great tit and zebra finch we calculated the change in GC content (ΔGC, see methods) per window due to insertions and due to deletions for both lineages. For the great tit, fixed deletions have acted to both increase GC content in some windows and decrease it in others, with the mean $\Delta GC = -0.88 \times 10^{-4}$, though median $\Delta GC = 0.22 \times 10^{-4}$. Insertions fixed in great tit windows however have largely acted to reduce GC content with mean $\Delta GC = -1.21 \times 10^{-4}$ and median $\Delta GC = -2.79 \times 10^{-4}$. ΔGC

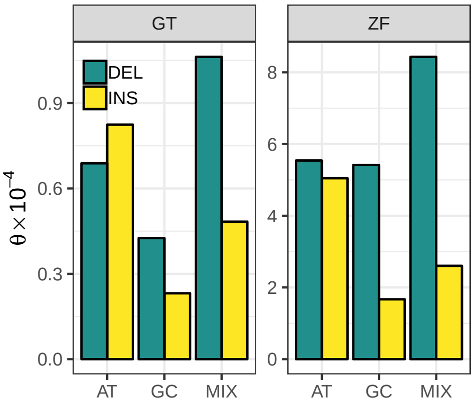


FIGURE 4.5: Estimated population scaled mutation rates ($\theta = 4N_e\mu$) for AT, GC and AT and GC mixed (MIX) insertions (INS) and deletions (DEL) in both the great tit (GT) and the zebra finch (ZF).

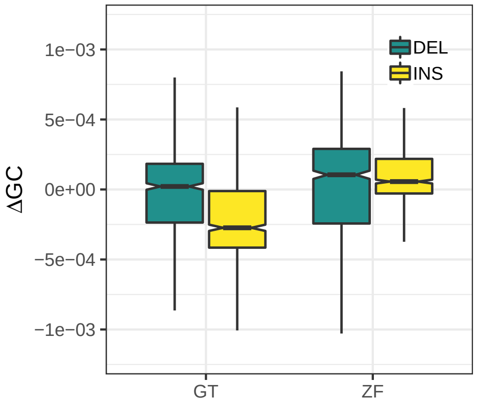


FIGURE 4.6: Change in GC content (ΔGC) on the branches leading to zebra finch (ZF) and great tit (GT) since their divergence, as a result of the fixation of small insertions (INS) and deletions (DEL).

for insertions is significantly lower than for deletions in the great tit (Wilcoxon rank sum, $W = 419554$, $p = 1.87 \times 10^{-21}$). In the zebra finch this picture differs with similar ΔGC for both insertions and deletions (Wilcoxon rank sum, $W = 310832$, $p = 0.056$), where for deletions mean $\Delta GC = -0.37 \times 10^{-4}$ and median $\Delta GC = 1.07 \times 10^{-4}$, and for insertions mean $\Delta GC = -1.55 \times 10^{-4}$ and median $\Delta GC = 0.54 \times 10^{-4}$.

When considering how ΔGC varies across the windows within our dataset, we see a positive correlation between ΔGC for insertions and ancestral GC content of the window, for both species, whilst for deletions we see the opposite, with negative relationships in both species (table 4.4, top panel figure 4.7). However, when we combine ΔGC for insertions and deletions to get an overall impact of small INDELs on GC content we see a different picture, with a slight positive correlation with ancestral GC content in the great tit, seemingly driven by elevated GC gain in the highest GC windows, with the bulk of windows demonstrating an overall GC reduction due to small INDELs (table 4.4, bottom panel figure 4.7). In the zebra finch, we instead report a negative correlation between ΔGC due to INDELs and ancestral GC content, generally with windows below mean GC content ($\bar{GC} = 0.4$, table 4.1) experiencing GC gains due to small INDELs and windows above mean GC experiencing GC reductions (table 4.4, bottom panel figure 4.7). We recapture these relationships when comparing ΔGC against recombination

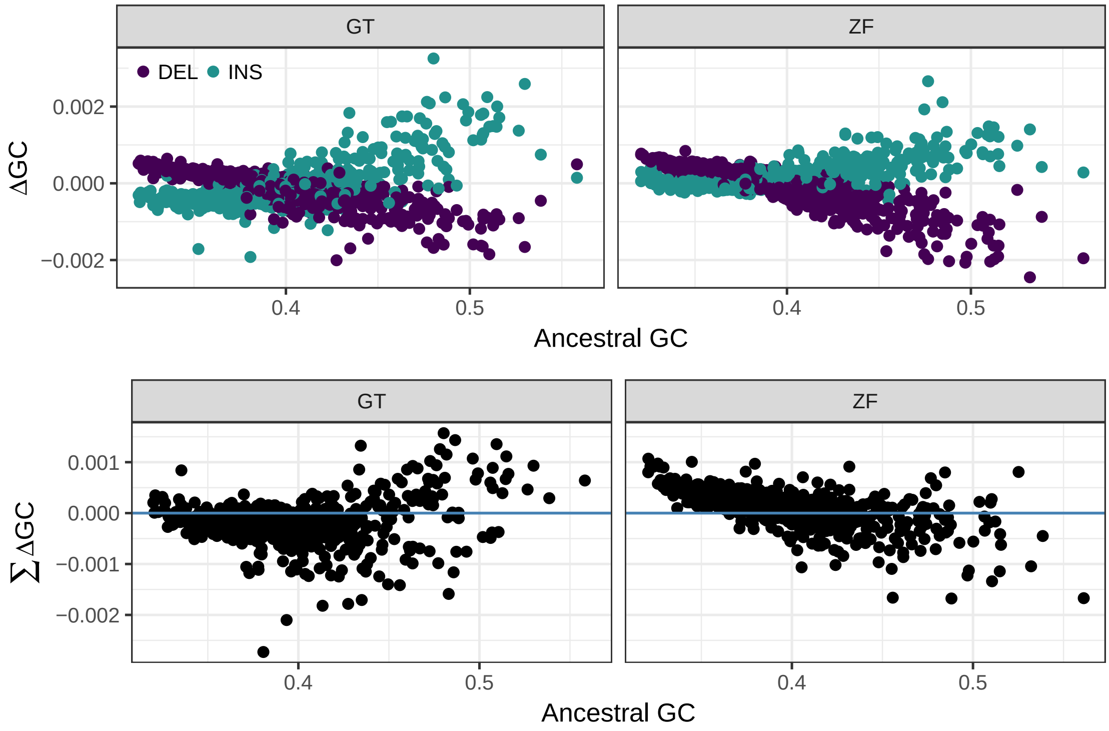


FIGURE 4.7: The relationship of ΔGC for insertions and deletions (top row) and delta GC for combined INDELs (∑ ΔGC, bottom row) against ancestral GC content in both great tit (GT) and zebra finch (ZF).

rate, and against B instead of ancestral GC content, other than for ΔGC for INDELs in the great tit where we only see negligible relationships (table C.2, figure C.10, figure C.11).

4.5 Discussion

Most contemporary studies addressing the role of GC biased gene conversion (gBGC) in genome evolution have focussed on its action within coding regions where it is confounded by the action of selection (Bolívar *et al.*, 2019; Corcoran *et al.*, 2017; Gossmann *et al.*, 2018; Ratnakumar *et al.*, 2010; Rousselle *et al.*, 2019) and processes like codon usage bias (Jackson *et al.*, 2017). Additionally few of these studies have looked at the impact of N_e on the strength of gBGC. Futhermore, research into the evolution of base composition has largely focussed on SNPs, influenced by gBGC and CpG hypermutability. To address these gaps, here we analyse 20 re-sequenced avian genomes from two species, the great tit (Corcoran *et al.*, 2017), and the zebra finch (Singhal *et al.*, 2015). Using a dataset of 1Mb orthologous windows we investigate the action of gBGC in the non-coding regions of these birds since their divergence, as well investigating the impact of small INDELs on base content evolution, a topic so far unaddressed.

4.5.1 Strength of gene conversion modulated by N_e

Our non-coding 1Mb orthologous window dataset yielded similar mean GC content (0.41, 0.41) and recombination rates (0.48, 0.41) in the great tit and zebra finch respectively. From our application of the Glémin *et al.* (2015) model to our dataset we obtained mean B estimates of 0.40 for the great tit and 0.90 for the zebra finch. Our estimate are similar to genome wide mean estimates of B in humans of 0.38 (Glémin *et al.*, 2015), falling at the lower end of the B range of 0.4 to 5 reported by Long *et al.* (2018) in a comparative study with taxa from across the tree of life. These B values are below 1 in both species, and thus at a level, particularly in the great tit, which may not be able to predominate over SW mutational biases. McVean and Charlesworth (1999) demonstrate that for synonymous codons, N_es much below 1 can lead to fixation of un-preferred over preferred codons when there is a mutational bias towards the un-preferred codon, here this parallels SW mutation biases and WS fixation biases due to gBGC. Additionally, the authors demonstrate that a greater proportion of un-preferred codons are fixed until $N_es \simeq 0.25$, and for $0.25 < N_es < 1$ selection for preferred codons can still be hampered by mutation favouring un-preferred codons. Thus it is likely gBGC in non-coding regions in these species is operating at reduced efficiency, particularly in the great tit.

When looking at variation in B within the genome we see significant correlations between B and recombination rate in both the great tit and zebra finch (table 4.2), consistent with gBGC being stronger in regions of higher recombination, as has been reported in humans (Glémin *et al.*, 2015) and as implied by correlations between GC content at 4-fold sites and recombination in flycatchers (Bolívar *et al.*, 2016). This relationship is stronger when using GC content as proxy for recombination, possibly as it is a better measure of long term recombination rate, but also likely as our recombination rate estimates are constrained by the density of the linkage maps available (Stapley *et al.*, 2008; van Oers *et al.*, 2014), whereas the calculated GC content is not. As local N_e also likely correlates with recombination rate these relationships were analysed with partial correlations using π as a confounding variable, precluding it as a driver for these trends.

When comparing between species we see a conserved biased gene conversion landscape, with per window B estimates correlated significantly between species (figure 4.3). In light of the strong correlations reported between both GC content and recombination rate between the species (correlation coefficients 0.83 and 0.72 respectively), this is perhaps unsurprising and likely a result of birds' conserved recombination hotspots (Singhal *et al.*, 2015), karyotype and synteny (Hansson *et al.*, 2010; Stapley *et al.*, 2008; van Oers *et al.*, 2014; Zhang *et al.*, 2014). However, whilst correlating strongly, B values are higher in zebra finch (figure 4.2a). As B is the product of b (the strength of biased gene conversion) and N_e, either parameter could be driving this increase. To separate the effects of these parameters we standardised our B estimates by π as a proxy for N_e. B/π estimates are similar in zebra finch and great tit whilst mean B is around twofold higher in the zebra finch (figure 4.2) consistent with the twofold larger N_e in this species (Corcoran *et al.*, 2017). This combined with our reported correlation between the ratio of the species' π estimates and the ratio of the species' B estimates supports the idea that the larger N_e in the zebra finch is driving the increased population scaled strength of gBGC in this species, and that by extension b has remained relatively stable since the species diverged. Additionally, this is in keeping with a number of previous studies. In birds, GC3 content has been reported to correlate with N_e (using life history traits as proxies) across the avian phylogeny (Weber *et al.*, 2014). Similarly, in great apes, B for 4-fold sites correlates with N_e (preprint: Bolívar *et al.*, 2018) and in rice species (*Oryza spp.*), selfers, which have reduced N_e, also have lower B estimates (Muyle *et al.*, 2011). However, analysis between more diverged species has failed to produce a relationship between B and N_e, with the authors suggesting that B only responds to N_e over small time-scales (Galtier *et al.*, 2018).

4.5.2 Non-coding equilibrium GC content

Our divergence based estimates of the mean non-coding equilibrium GC content, GC$*_{div}$, are lower than the current GC content in both the great tit (0.39 versus 0.40) and the zebra finch (0.38 versus 0.40), and this trend extends to the majority of windows in the dataset (figure C.7). This finding is at odds with previous avian studies reporting that current GC content is below GC$*$ in most avian lineages (Bolívar *et al.*, 2016; Rousselle *et al.*, 2019; Weber *et al.*, 2014). However, these studies are largely focussed on GC3 content in coding regions. In birds, coding regions have been seen to have higher GC content (Weber *et al.*, 2014) and higher recombination rates (Singhal *et al.*, 2015) possibly as a result of higher gene density on micro-chromosomes which have elevated recombination rates and GC content (Burt, 2002; Stapley *et al.*, 2008; van Oers *et al.*, 2014). Indeed this is supported in our dataset where non-coding GC content is $\sim$ 10% lower than coding sequence GC content (table C.3).

Our B estimates correlate strongly with GC$*_{div}$ as well as with the distance from GC equilibrium (GC$*_{div}$ - GC) suggesting that areas of high B have elevated GC$*_{div}$ and are further from equilibrium. This is consistent with correlations between GC$*$ and recombination rate previously reported in zebra finch and long-tailed finch (*Poephila acuticauda*) (Singhal *et al.*, 2015), as well as in rice species (Muyle *et al.*, 2011). Furthermore, both recombination rates and B estimates are highest in windows that are still increasing towards GC$*_{div}$ and lower in regions close to, or decreasing towards GC$*_{div}$ in both species (figure C.9). Taken together these results suggest that non-coding GC content has been decreasing towards a lower GC$*_{div}$ in low recombining regions since the great tit zebra finch split. This may be a result of a decreased efficacy of gBGC, with median B estimates below 0.5 in regions above GC$*_{div}$ (figure C.9a), reducing the proportion of GC alleles fixing and allowing for more AT biased fixation patterns (see McVean and Charlesworth, 1999) and a slight erosion of ancestral GC levels, possibly stemming from historically lower N_e, with both species showing evidence of recent population increases (Balakrishnan and Edwards, 2008; Corcoran *et al.*, 2017; Laine *et al.*, 2016). Additionally, with lower recombination rates in non-coding regions than in coding regions (Singhal *et al.*, 2015) it also raises the possibility that non-coding and coding

GC content have been diverging in these species, such as may also be occurring in some species of rice where non-coding GC content is above equilibrium when GC3 is below equilibrium (Muyle *et al.*, 2011).

Our estimates of GC$*_{pol}$ derived from our application of the Glémin *et al.* (2015) model applied to the polymorphism dataset paints a different picture. The mean GC$*_{pol}$ estimates from this approach are much higher, 0.63 versus 0.39 for great tit and 0.72 versus 0.38 for zebra finch. The two approaches differ in the time frame they analyse, with the divergence approach estimating a more long term GC$*$ spanning the entire branch length from the great tit - zebra finch split, where as the polymorphism approach provides more recent information only going as far back as $2N_e$ generations. As such, comparing the two approaches can provide an indication of how B has changed towards the present. Our elevated estimates for the polymorphism based GC$*_{pol}$ suggests an increase in B over the past $2N_e$ generations. As recombination rates are relatively stable and conserved in these species it seems likely this is a result of increased N_e due to a population expansion as has been previously reported for the great tit (Corcoran *et al.*, 2017; Laine *et al.*, 2016) and the zebra finch (Balakrishnan and Edwards, 2008; Corcoran *et al.*, 2017). Additionally the increase in GC$*$ is greater for the zebra finch which also shows evidence of a larger population growth (Corcoran *et al.*, 2017). Interestingly, the converse has been reported in *Drosophila melanogaster*, where longer term estimates of B are higher than those from the Glémin *et al.* (2015) model, suggesting a reduction in B over time (Jackson *et al.*, 2017). Our GC$*$ estimates from these two methods suggest that increases in N_e have shifted non-coding base composition from moving towards reduced GC content to increased GC content in these birds. Although, this is likely an oversimplified picture with N_e likely fluctuating, and consequently B, since the great tit and the zebra finch split $\sim$ 40 million years ago (Barker *et al.*, 2004).

4.5.3 Small INDELs contribute to GC content change

Our θ and π estimates for INDELs in different genomic contexts were higher for the zebra finch than the great tit, consistent with a larger N_e in this species. We see a marked deletion bias for INDELs containing only GC bases and for INDELs containing a mix of AT and GC bases (figure 4.5) consistent with previously reported deletion biases in a wide range of species (Hu *et al.*, 2011; Keightley *et al.*, 2009; Kvikstad and Duret, 2014; Nam and Ellegren, 2012; Presgraves, 2006; Taylor *et al.*, 2004). However, for AT containing INDELs this relationship disappears, with a rDI (ratio of insertions to deletions) of 1.1 in zebra finch and 0.84 in the great tit; an insertion bias. Whilst striking, this difference is in line with previously reported higher rDIs in GC rich introns in humans (Wang and Yu, 2011). Through polymerase slippage it is energetically more expensive to create an insertion than a deletion as it requires a previously duplicated strand of DNA to denature and be re-replicated along the whole length of an insertion, but not for deletions (Petrov, 2002b). Additionally, sequence melting temperature reduces with increase AT content, making AT rich regions less stable (Fryxell and Zuckerkandl, 2000). As such, it may be that in AT rich regions, where INDELs are more likely AT rich (as seen in our alignment dataset: figure C.12), the energetic cost of insertion formation is reduced relative to deletion formation, allowing for an erosion of the deletion bias.

When we consider the selective pressures acting on these INDELs, our analysis provides evidence for selection or a selection like force favouring insertions and disfavouring deletions for GC containing and mixed non-coding INDELs in both species (table C.1). The more negative γ estimates for deletions relative to insertions is consistent with previous reported estimates in the great tit (Barton and Zeng, 2019) and *D. melanogaster* (Barton and Zeng, 2018). However, as with the mutation rate estimates, γ for AT containing INDELs shows a different relationship, where deletions show more positive estimates than insertions, although in the zebra finch both seem to be weakly selected for where as in the great tit they are weakly selected against (table C.1).

The fixation of small INDELs since the common ancestor of the zebra finch and great tit has on net acted as a GC decreasing force in the great tit, and an increasing one in the zebra finch (figure 4.6, figure 4.7). When considering the per window change in GC content (ΔGC) due to insertions and deletions separately we see that as ancestral GC content increases, intuitively so does the amount of GC lost and gained through small INDELs. If the GC content of INDELs is above the window GC content (figure C.12), then this results in deletions reducing GC content and insertions increasing it. This indeed appears to be the case in windows with above 40% GC content, in both species (figure 4.7, figure C.12), though the reason behind this is unclear.

4.5.4 Conclusion

In summary we show marked conservation of the underlying strength of gBGC, b in the zebra finch and great tit, with the population scaled strength of gBGC, B, larger in the zebra finch in proportion to the species' larger N_e. Additionally, B seems weak ($B < 1$) in the majority of our non-coding data, and non-coding GC has seemingly been decreasing since the two species diverged, contrary to previous work on coding regions. Furthermore, we demonstrate that neither polymorphic or fixed INDELs are GC conservative in nature and have the ability to shape genomic base composition, it would be interesting to incorporate this into future modelling frameworks addressing GC content evolution, although this is non-trivial.

4.6 Supplementary Material

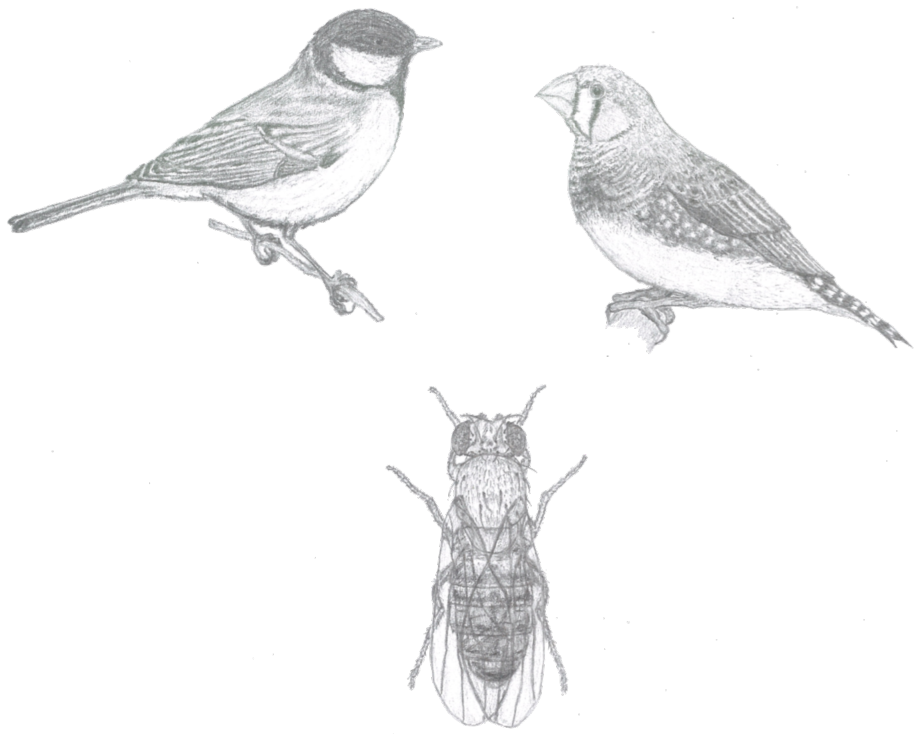
Supplementary figures and tables are available in Appendix C.

4.7 Acknowledgements

We thank Pádraic Corcoran for an initial implementation of the window pipeline. This work was supported by a PhD studentship funded by the Department of Animal and Plant Sciences, University of Sheffield, to H.J.B. Support was also provided by the Natural Environment Research Council via a research grant awarded to K.Z. (NE/L005328/1).

Chapter 5

General conclusion and discussion



5.1 Selection and small INDELs

Insertions and deletions (INDELs) have not received the same level of attention as single nucleotide polymorphisms (SNP), leaving them understudied in comparison. Whilst their mutation rate is markedly lower than that of SNPs (Barton and Zeng, 2019; Montgomery *et al.*, 2013), they nonetheless contribute new genetic variation on which selection can act. Arguably they may even contribute more to divergence when their length is considered (Britten, 2002; Britten *et al.*, 2003). Thus, understanding the selective forces operating on INDELs is of great interest. However, to date, there has been little attempt to directly quantify the strength of selection acting on them, largely due to a lack of methods available and the challenges in doing so. One such challenge is the need to separate called INDELs into insertions and deletions, which requires determining if the long or short variant is ancestral. However, this process is error prone, and when combined with a mutational bias towards deletions over insertions (Besenbacher *et al.*, 2015; Keightley *et al.*, 2009; Schrider *et al.*, 2013; Yang *et al.*, 2015) could cause spurious signatures of selection on insertions (Barton and Zeng, 2018; Hernandez *et al.*, 2007; Kvikstad and Duret, 2014).

In this thesis I apply a novel model described in Chapter 2 to estimate the distribution of fitness effects (DFE) for small INDELs ($\leq 50bp$) in coding regions in *Drosophila melanogaster* (Chapter 2) and the great tit (*Parus major*) (Chapter 3). Additionally, I characterise the INDEL DFE in non-coding regions in the great tit (Chapter 3) and the zebra finch (*Taeniopygia guttata*) (Chapter 4). The model estimates the population scaled mutation rate ($\theta = 4N_e\mu$), selection coefficient ($\gamma = 4N_es$) and the rate of ancestral misidentification (ϵ), it controls for this error and for demography using a set of putatively neutral sites as reference as in Eyre-Walker *et al.* (2006). In Chapter 2 I demonstrate that the model performs well, providing accurate estimates across a wide range of parameters.

Through application of the model to a *D. melanogaster* resequencing dataset (Pool *et al.*, 2012), I demonstrate that the DFE for INDELs in coding regions is bimodal, characterised by a class of strongly deleterious INDELs, accounting for the majority of coding

INDEL events, and a class of weakly selected sites accounting for the remaining minority of INDELs. Of the weakly selected variants, deletions are more deleterious with more negative γ estimates. There is also some evidence for weak positive selection operating on insertions, however this result is dependant on the model used. The best fit DFE for coding INDELs in the great tit obtained from a high coverage great tit resequencing dataset (Corcoran *et al.*, 2017) is markedly similar, again being best explained by a bimodal distribution. Consistent with the *D. melanogaster* DFE the majority (96%) of coding INDELs are strongly deleterious with the remainder weakly selected against, again with deletions in this class having more negative γ estimates than insertions. These results are consistent with the more deleterious nature of deletions as theorised due to their larger base impact (Petrov, 2002b; Sjödin *et al.*, 2010) and as previously inferred from allele frequencies and divergence levels (Chintalapati *et al.*, 2017; Leushkin and Bazykin, 2013; Presgraves, 2006; Sjödin *et al.*, 2010). In *D. melanogaster* we also demonstrate that INDELs that shift the reading frame (i.e. those not a multiple of 3 in length) have more negative γ estimates, a finding that is also reflected in the INDEL length distribution in the great tit, with in-frame INDELs more numerous. Additionally, there is a reduction in the proportion of frame-shifting INDELs in more conserved genes.

I extended the coding region analysis beyond the polymorphism data and incorporated divergence data in order to estimate the proportion of fixations driven by positive selection (α) for insertions and deletions. In the *D. melanogaster* dataset this yielded estimates of 100% for insertions and 82% for deletions. In the great tit α for insertions is 71% and for deletions 86%. The higher value for deletions is consistent with them being more deleterious than insertions, and thus having lower fixation probabilities, yielding higher estimates of α. Whilst the *Drosophila* data shows the opposite trend with α at 100%, this is a result of the best fit bimodal model having positive γ values for the weakly selected class of deletions, which can not be used to estimate a neutral fixation probability, leaving only the strongly deleterious class to be used in the calculation, but they are too deleterious to contribute to divergence. The truth is likely more complex, with some effectively neutral insertions present in the *Drosophila* data, but we lack power with only 17 haplotypes (discussed in more detail in section 5.3). The two species deletion α estimates are however, very similar.

My analysis of non-coding INDELs was largely confined to the great tit dataset. Here we see that the DFE is best described by a gamma distribution. In non-coding regions $\sim 80\%$ of insertions and $\sim 52\%$ of deletions are effectively neutral, with γ estimates between 0 and 1. The remaining INDELs are evenly split between the other, more deleterious, selective categories ($-1 > \gamma > -10$, $-10 > \gamma > -100$ and $\gamma < -100$). In Chapter 4, I apply a one class model to both the great tit and the zebra finch data, in both species the majority of deletions are more deleterious than insertions. However, in this analysis we see positive γ estimates for insertions in a number of models, similar to the coding DFEs for great tit and *D. melanogaster*. If these positive non-coding γ estimates are a signature of a neutral 'selection like' process such as insertion biased gene conversion (Leushkin and Bazykin, 2013), it is conceivable that this would not have been detected in the Chapter 3 analysis which uses INDELs in ancestral repeats as neutral reference. This reference would also have been subject to such a force, obscuring its signature, whereas through the use of weak to weak and strong to strong SNPs as neutral reference, which are not influenced by any form of gene conversion, it can be detected. In the great tit I also demonstrated that non-coding INDEL diversity is reduced near exons and in areas of low recombination as a result of linked selection, though how much can be attributed to purifying selection versus positive selection is unresolved. Additionally, I present some evidence for the mutagenic effect of recombination on INDEL diversity and stronger selection against longer INDELs, particularly insertions.

5.2 GC biased gene conversion and INDELs' contribution to base composition

Base composition and its determinants are central to understanding how genomes evolve (Eyre-Walker and Hurst, 2001). Base composition is largely shaped by mutational biases, generally biased towards AT mutations, and fixation biases of point mutations. The latter can be distorted by selection, such as selection on codon usage (Duret, 2002; Sharp *et al.*, 1995) or 'selection like' processes such as GC biased gene conversion (gBGC). Additionally, base composition has the potential to be impacted by the fixation of small INDELs. Most work to date on gBGC has focussed on its role in coding regions, and a large portion of that work with a view of understanding how it confounds signatures of selection, rather than studying gBGC itself. Additionally, there is little work to date addressing the role of small INDELs. In this thesis I address these themes through the application of the Glémin *et al.* (2015) model to estimate the strength of GC biased gene conversion ($B = 4N_eb$) in the great tit and in the zebra finch, across orthologous non-coding windows in the two species. Furthermore, I apply the model from Chapter 2, to characterise INDELs of different base composition and assess their contribution to genomic base composition.

The dataset of orthologous windows in the zebra finch and great tit demonstrated a remarkably conserved recombination and GC content landscape between the species, consistent with birds conserved karyotype and synteny (Stapley *et al.*, 2008; van Oers *et al.*, 2014) and stable recombination hotspots (Singhal *et al.*, 2015). Concordantly, B estimates are also correlated in the two species, although the mean magnitude is around twofold higher in the zebra finch ($\bar{B} = 0.9$) than the great tit ($\bar{B} = 0.4$), consistent with its twofold greater effective population size (Corcoran *et al.*, 2017). This suggests that the larger N_e is the driver for its larger B and the underlying rate of gene conversion (b) is probably similar between the species, this is consistent with studies in great apes (preprint: Borges *et al.*, 2018), birds (Weber *et al.*, 2014) and rice (Muyle *et al.*, 2011), which also provide support for the role of N_e modulating the strength of gBGC, although it has been suggested that this only holds over short evolutionary distances (Galtier *et al.*, 2018).

Using this dataset I also estimated the per window equilibrium GC content (GC$*$), using both substitution rates (yielding GC$*_{div}$) and the fixation rates inferred from the θ and γ estimates obtained for the polymorphism dataset (yielding GC$*_{pol}$) . These two methods reflect different times scales, with the divergence based approach covering everything from the species split to recent fixations, whilst the polymorphism based approach goes only as far back as the past $2N_e$ generations so is more contemporary. Mean GC$*_{div}$ estimates are lower than current GC levels, suggesting that the GC content of these birds has been decreasing, contrary to reports of GC dynamics in coding regions in birds (Bolívar *et al.*, 2016; Rousselle *et al.*, 2019; Weber *et al.*, 2014). Our mean GC$*_{pol}$

estimates however are much higher, and above current GC levels. This is consistent with previously reported evidence for population expansions in these birds (Balakrishnan and Edwards, 2008; Corcoran *et al.*, 2017; Laine *et al.*, 2016), which may have increased the efficacy of GC biased gene conversion and thus increased GC$*$. Additionally as our mean estimates of B which reflect the recent strength of gBGC are relatively low ($\bar{B} < 1$), it stands to reason that historical B is likely to have been even lower with smaller N_e, thus allowing for the erosion of the non-coding GC content through inefficient gBGC.

Finally, I analysed small INDELs contribution to base composition in these two species. My θ estimates for GC INDELs and INDELs containing a mix of AT and GC bases support the deletion bias trend seen in previous chapters and elsewhere (Besenbacher *et al.*, 2015; Keightley *et al.*, 2009; Schrider *et al.*, 2013; Yang *et al.*, 2015). However, for solely AT containing INDELs there is not a marked deletion bias. I suggest this may be a result of the higher energetic cost of insertions over deletions (Petrov, 2002b) being reduced in less thermodynamically stable AT rich regions. My γ estimates for GC and mixed insertions are consistently positive in both species whereas those for deletions are negative, this is consistent with previous work suggesting the more deleterious nature of deletions (Petrov, 2002b; Sjödin *et al.*, 2010), and the positive values for insertions may reflect the action of a force such as insertion biased gene conversion (Leushkin and Bazykin, 2013). I analysed the INDELs fixed on each lineage since the divergence of the great tit and zebra finch, showing they have acted to reduced GC content in the great tit and increase it in the zebra finch.

5.3 Current limitations

Throughout this thesis a number of methodological limitations have emerged, the predominant of which are statistical power, the importance of neutral reference choice and calling constraints on INDEL length.

Firstly, statistical power. For the analyses of INDELs in coding regions in Chapters 2 and 3, statistical power is an issue. As INDELs in genes are generally strongly deleterious the majority are extremely rare and thus not segregating in our relatively small samples

(17 haplotypes for *D. melanogaster*, 20 haplotypes for the great tit). As a result we lack information on the deleterious end of the DFE. Equally as only the minority of coding INDEL events ($\sim 4\%$ according to my Chapter 3 analysis) are segregating in the samples, sub-setting the coding data is problematic, so making comparisons based on INDEL length, or genomic region, or on the level of individual genes or groups of genes is unrealistic without much larger sample sizes. Additionally it raises the possibility that the true best fit DFEs may differ from those reported.

Secondly, choice of neutral reference. Deciding on a neutral reference for INDEL analyses is non trivial, with fewer/no suitable sites within coding regions like SNPs (i.e. fourfold degenerate sites). However, non-coding regions likely contain functionally important and conserved sites, and there is some evidence for selection on INDEL length in introns (Ometto *et al.*, 2005), making them problematic. In Chapter 2 I estimate the INDEL DFE separately using non-coding sites and fourfold SNPs as neutral reference. However to use the SNP reference and still be able to estimate the strongly deleterious end of the DFE is only possible due to the existence of previous mutation experiments (Schrider *et al.*, 2013) to determine the SNP-INDEL mutation rate ratio in *D. melanogaster*. Obviously, the same depth of literature is not readily available in less studied organisms. In Chapter 3 I instead use ancestral repeat elements in the great tit which should not be under any selective constraint, however INDEL rates can be higher in repetitive sequence contexts (Ananda *et al.*, 2013; Montgomery *et al.*, 2013), thus this choice is also not perfect, however similar results are obtained when using non-coding INDELs as neutral reference. Using a neutral reference with selective sites within it can impact the resulting DFE, this is discussed in detail in Chapters 2 and 3, briefly, simulation results show the presence of selected sites in the neutral reference can lead to an underestimation of purifying selection and an overestimated fixation rate which can result in an underestimate α, but this depends on how strong the selection is on the neutral reference relative to the focal sites.

Finally INDEL length. In this thesis I have restricted my analysis to 'small' INDELs, here $\leq 50bp$, a somewhat arbitrary cut off, but necessary due to the poor ability to call longer variants reliably from short read data. Although dedicated packages exist for calling large INDELs and structural variants from this type of data, when multiple packages were applied to the great tit dataset they showed remarkably little overlap in called variants (P. Corcoran personal communication). As a result our analyses are not a full representation of INDELs role in the genome. Additionally, it makes addressing questions relating to genome length problematic, as demonstrated by Petrov (2002a) and Gregory (2003, 2004).

5.4 Concluding remarks

This thesis has advanced our knowledge of the selective landscape of small insertions and deletions in both coding and non-coding regions, providing evidence for strong purifying selection in coding regions, weak purifying, positive and linked selection in non-coding regions, as well as demonstrating that INDELs impact the GC content of two passerines. Additionally it has added to the small body of literature quantifying GC biased gene conversion in non-coding regions.

Moving forward, applying the models used in this thesis to a much larger sample of haplotypes, such as the 1000 genomes project (1000 Genomes Project Consortium, 2010), would allow for a more informative view of the INDEL DFE, providing more power to generate a higher resolution DFE. Additionally, it would be interesting to supplement the growing body of avian resequencing data with longer read data, such as from PacBio sequencing, to enable higher accuracy calling of larger INDELs and to generate a more complete view of INDEL dynamics across a broader, more speciose phylogenetic sample.

Appendix A

Supplementary Material to Chapter 2

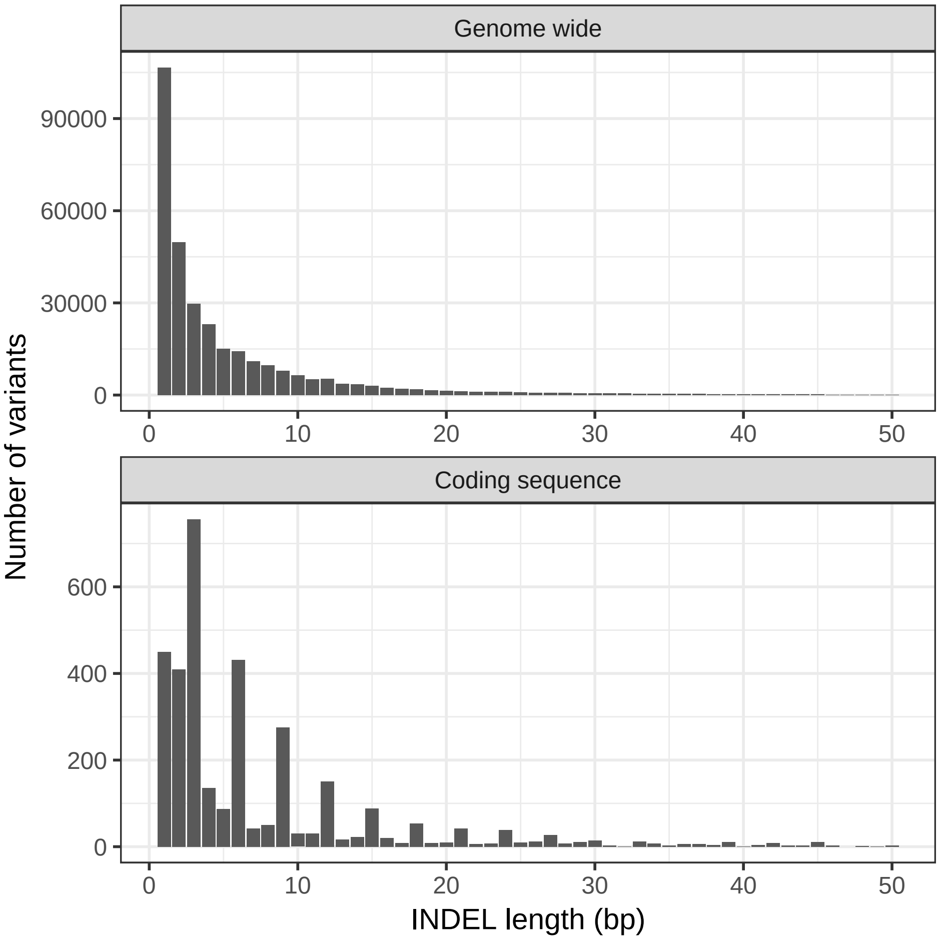


Figure A.1: Length distribution of INDEL polymorphism in *D. melanogaster*. Upper panel: the distribution of all INDELs called ($\leq 50bp$). Lower panel: the distribution of INDELs within CDS regions.

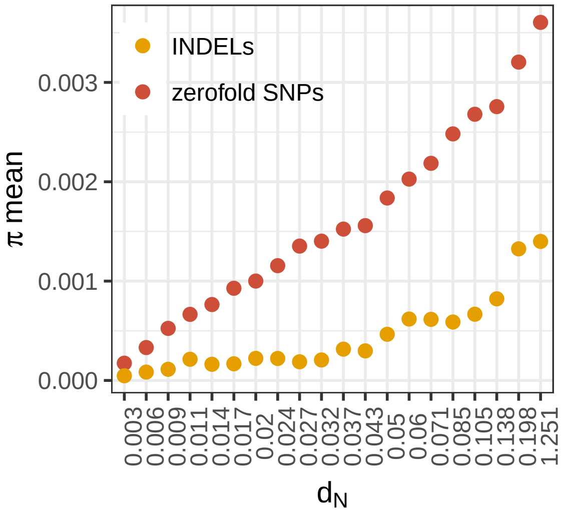


FIGURE A.2: The correlation between d_N and π_0 or π_{INDEL}. The genes were binned based on dN into 20 equal-sized groups.

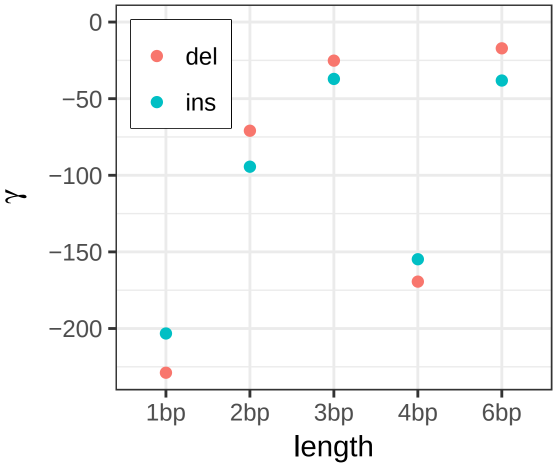


FIGURE A.3: Estimates of γ for polymorphic INDELs of different lengths. The INDELs in the *D. melanogaster* dataset were divided into the following length categories: 1bp, 2bp, 3bp, frameshifting (≤4bp), non-frameshifting (≤6bp). Non-coding INDELs with the same lengths were used as the neutral reference. The data were analysed using a model with 1 class of sites in the selected region and uniform mutation rate across the selected and neutral regions.

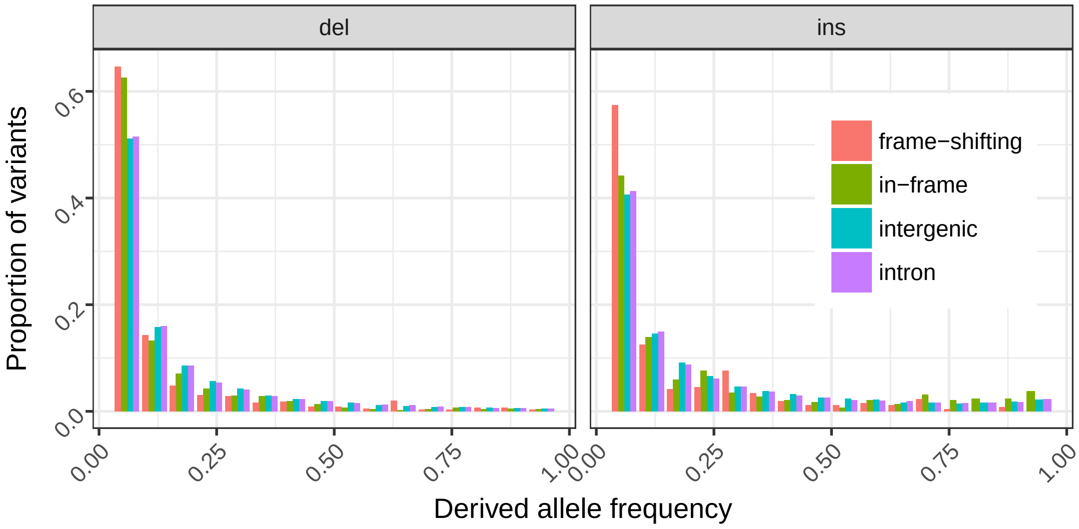


FIGURE A.4: The site-frequency spectra for insertions and deletions in different genomic regions in the *D. melanogaster* dataset.

Table A.1: Results based on fitting the new models to the INDELs within protein-coding regions from *D. melanogaster*. Non-coding INDELs were used as the neutral reference, and the mutation rates to insertions and deletion were assumed to be equal between selected and neutral regions.

Model	Parameters for INDELs in the CDS regions									ΔAIC
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}			0
$C = 2$	MLE	1.8×10^{-5}	1.98	0.023	5.3×10^{-5}	-1.69	0.016			
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}			
	MLE	7.2×10^{-4}	-1566.4	3.6×10^{-5}	0.0011	-642.5	1.6×10^{-5}			
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}			9
$C = 3$	MLE	1.86×10^{-5}	1.41	0.022	5.5×10^{-5}	-3.03	0.0040			
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}			
	MLE	0.0002	-657.5	0.0003	0.0011	-771.9	2.4×10^{-5}			
	Name	θ_3^{ins}	γ_3^{ins}	ϵ_3^{ins}	θ_3^{del}	γ_3^{del}	ϵ_3^{del}			
	MLE	0.0005	-4584.3	1.5×10^{-5}	5.1×10^{-6}	885.9	0.169			
Continuous	Name	θ^{ins}	α^{ins}	b^{ins}	ϵ^{ins}	θ^{del}	α^{del}	b^{del}	ϵ^{del}	484
	MLE	0.0007	0.71	2596.7	0.033	0.0012	0.63	2417.1	0.038	
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}			6831
$C = 1$	MLE	0.0007	-96.3	0.096	0.0011	-70.0	0.054			

Table A.2: The effects of the presence of selected variants in the neutral reference on the estimation of the strength of selection on variants in the selected dataset. The sample size is 17 in all cases, the mutation rate is uniform across the genome, and the polarisation error rate is zero. The neutral reference dataset was generated with $\gamma = -3.5$. The selected dataset was generated using three different γ values. The simulated data were analysed by a model that considers a single class of selected sites, involves the use of the r parameters (see Eq. (11)), and assumes uniform mutation rate. The results are based on 50 replicates. $\bar{\mu}$ is the average fixation probability.

γ for selected variants		$\bar{\mu}$	
True	Mean (MLEs)	True	Mean (MLEs)
-1.5	6.997	0.431	7.004
-3.5	-0.018	0.109	0.992
-10	-7.268	4.5×10^{-4}	0.005

TABLE A.3: Simulations showing the effects of strong purifying selection parameter estimations when the neutral and selected regions are allowed to have separate mutation rate parameters. The sample size is 17, and the results are based on 50 replicates. The data were simulated by assuming that the DFE follows a reflected gamma distribution with different shape (a) and scale (b) parameters. Note that the mean for γ is $\bar{\gamma} = ab$. θ is the scaled mutation rate per site. $\bar{\mu}$ is the average fixation probability.

$\bar{\gamma}$		a		b		θ		$\bar{\mu}$	
True	Mean (MLEs)	True	Mean (MLEs)	True	Mean (MLEs)	True	Mean (MLEs)	True	Mean (MLEs)
-1000	-193.4	0.3	0.33	3333.3	586.1	2.12×10^{-4}	1.31×10^{-4}	0.1035	0.1748
-5000	-309.2	0.3	0.35	16666.7	883.3	2.12×10^{-4}	8.86×10^{-5}	0.0639	0.1609

TABLE A.4: Results based on fitting the new models to the INDELs within protein-coding regions from *D. melanogaster*. SNPs from 4-fold sites were used as the neutral reference. But the neutral and the selected regions were assumed to have their separate mutation parameters.

Model	Parameters for INDELs in the CDS regions									ΔAIC
Continuous	Name	θ^{ins}	α^{ins}	b^{ins}	ϵ^{ins}	θ^{del}	α^{del}	b^{del}	ϵ^{del}	0
	MLE	8.2×10^{-5}	0.36	968.5	0.0	6.6×10^{-4}	0.57	2195.8	0.0037	
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}			6
$C = 2$	MLE	1.5×10^{-5}	-1.05	0.0032	3.9×10^{-5}	-3.08	0.0120			
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}			
	MLE	6.6×10^{-5}	-80.2	2.0×10^{-6}	1.8×10^{-4}	-56.2	2.1×10^{-6}			
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}			16
$C = 3$	MLE	1.6×10^{-5}	-2.46	0.032	2.4×10^{-4}	-88.4	0.0080			
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}			
	MLE	1.5×10^{-6}	15.8	0.0002	1.5×10^{-6}	999.9	0.4025			
	Name	θ_3^{ins}	γ_3^{ins}	ϵ_3^{ins}	θ_3^{del}	γ_3^{del}	ϵ_3^{del}			
	MLE	9.5×10^{-5}	-138.0	1.4×10^{-6}	4.4×10^{-5}	-3.77	0.0012			
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}			67
$C = 1$	MLE	2.6×10^{-5}	-2.83	0.039	9.0×10^{-5}	-6.16	0.011			

TABLE A.5: Results based on fitting the new models to the INDELs within protein-coding regions from *D. melanogaster*. SNPs from 4-fold sites were used as the neutral reference. The mutation rate ratio between SNPs and INDELs, and that between deletions and insertions, were fixed to 12.2 and 5, respectively (Schrider *et al.*, 2013). Note that only results based on discrete models with C = 2 or 3 classes of selected sites are presented. This is because analyses using either a discrete model with C = 1 class of sites or a continuous model with γ following a reflected gamma distribution failed to converge to biologically meaningful regions of the parameter space. This is probably due to these models are highly unrealistic given the constraints on the mutation rates.

Model	Parameters for INDELs in the CDS regions							ΔAIC
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	0
$C = 2$	MLE	1.9×10^{-4}	-284.1	1.2×10^{-4}	0.0010	-454.8	6.2×10^{-5}	
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	MLE	1.6×10^{-5}	-1.31	0.0092	4.9×10^{-5}	-3.77	0.0082	
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	9
$C = 3$	MLE	2.1×10^{-5}	-408.1	3.8×10^{-4}	0.0010	-1009.3	0.0195	
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	MLE	1.7×10^{-5}	-3.25	0.0154	3.9×10^{-5}	-3.03	0.0081	
	Name	θ_3^{ins}	γ_3^{ins}	ϵ_3^{ins}	θ_3^{del}	γ_3^{del}	ϵ_3^{del}	
	MLE	2.8×10^{-6}	13.15	0.0994	8.1×10^{-5}	-36.11	0.0096	

Table A.6: Estimates of the mutation rate ratios between SNPs and INDELs and between deletions and insertions.

Paper	SNP/INDEL	deletions/insertions
Petrov & Hartl (1998, Mol Biol Evol 15:293-302)	6.9	8.7
Haag-Liautard et al. (2007, Nature 445:82-85)	4.2	3.0
Schrider et al. (2013, Genetics 194:937-953)	12.2	5.0

Table A.7: Results based on fitting the new models to the INDELs within protein-coding regions from *D. melanogaster*. SNPs from 4-fold sites were used as the neutral reference. The mutation rate ratio between SNPs and INDELs, and that between deletions and insertions, were fixed to 6.9 and 8.7, respectively (Petrov and Hartl, 1998). Note that only results based on discrete models with C = 2 or 3 classes of selected sites are presented. This is because analyses using either a discrete model with C = 1 class of sites or a continuous model with γ following a reflected gamma distribution failed to converge to biologically meaningful regions of the parameter space. This is probably due to these models are highly unrealistic given the constraints on the mutation rates. The likelihood surface for the C = 3 model also appears to be somewhat flat a parameter combination without large γ values only has a slightly lower log likelihood (516951.999 versus 516951.842).

Model	Parameters for INDELs in the CDS regions							ΔAIC
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	0
$C = 2$	MLE	2.1×10^{-4}	-310.0	7.3×10^{-5}	0.0019	-909.7	1.0×10^{-4}	
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	MLE	1.6×10^{-5}	-1.29	0.0094	4.9×10^{-5}	-3.81	0.0083	
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	8
$C = 3$	MLE	1.6×10^{-5}	-2.93	5.4×10^{-4}	1.3×10^{-6}	1000	0.0012	
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	MLE	2.3×10^{-4}	-394.7	3.5×10^{-4}	5.2×10^{-5}	-4.61	1.3×10^{-4}	
	Name	θ_3^{ins}	γ_3^{ins}	ϵ_3^{ins}	θ_3^{del}	γ_3^{del}	ϵ_3^{del}	
	MLE	2.9×10^{-6}	1000	0.1473	0.0021	-1046.9	0.0292	

Table A.8: Results based on fitting the new models to the INDELs within protein-coding regions from *D. melanogaster*. SNPs from 4-fold sites were used as the neutral reference. The mutation rate ratio between SNPs and INDELs, and that between deletions and insertions, were fixed to 4.2 and 3, respectively (Haag-Liautard *et al.*, 2007). Note that only results based on discrete models with C = 2 or 3 classes of selected sites are presented. This is because analyses using either a discrete model with C = 1 class of sites or a continuous model with γ following a reflected gamma distribution failed to converge to biologically meaningful regions of the parameter space. This is probably due to these models are highly unrealistic given the constraints on the mutation rates.

Model	Parameters for INDELs in the CDS regions							ΔAIC
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	0
$C = 2$	MLE	8.9×10^{-4}	-1451.6	4.1×10^{-5}	0.0027	-1292.7	3.0×10^{-5}	
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	MLE	1.7×10^{-5}	-1.45	0.0104	5.0×10^{-5}	-3.89	0.0083	
Discrete	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	8
$C = 3$	MLE	1.6×10^{-5}	-2.72	0.0046	3.8×10^{-5}	-2.98	0.0065	
	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	MLE	2.4×10^{-6}	13.69	0.14	0.0028	-5548.2	0.0288	
	Name	θ_3^{ins}	γ_3^{ins}	ϵ_3^{ins}	θ_3^{del}	γ_3^{del}	ϵ_3^{del}	
	MLE	9.5×10^{-4}	-1763.1	3.4×10^{-5}	1.3×10^{-4}	-43.92	0.0045	

Table A.9: A comparison of parameter estimates obtained from analyses based on different mutation rate ratio estimates.

Source of mutation rate ratios	Weakly selected sites		α (%)		
	γ^{ins}	γ^{del}	INDELs	insertions	deletions
Petrov and Hartl (1998)	-1.29	-3.81	71.6	59.5	81.6
Haag-Liautard *et al.* (2007)	-1.45	-3.89	72.9	61.5	82.3
Schrider *et al.* (2013)	-1.31	-3.77	71.5	59.7	81.3

Appendix B

Supplementary Material to Chapter 3

Table B.1: Estimates of polarisation error used to correct the site frequency spectrum prior to calculating the summary statistics. CDS estimates are from the best fit model in table B.2, from the weakly deleterious site class, the ancestral repeat (AR) estimates are also from this model. Non-coding estimates are from the best fit model in table B.4.

Region	ϵ_{ins}	ϵ_{del}
CDS	0.0799	0.0368
Non-coding	0.0110	0.0166
AR	0.0302	0.0261

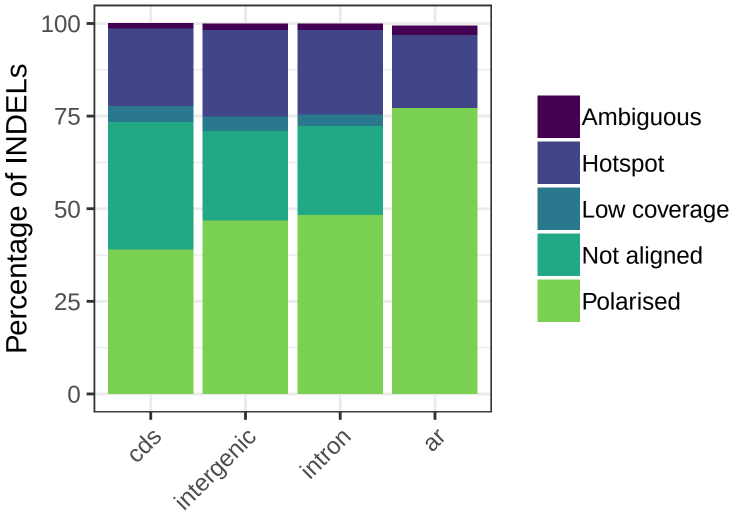


Figure B.1: Proportion of INDELs that were polarised in different genomic regions. Ambiguous refers to sites where both great tit alleles are represented in the outgroups, for example due to ancestral polymorphism. Note that due to the way ancestral repeats (ar) were identified from the genome alignment, it is not possible for INDELs to fail polarisation due to being not aligned or have 'low coverage'.

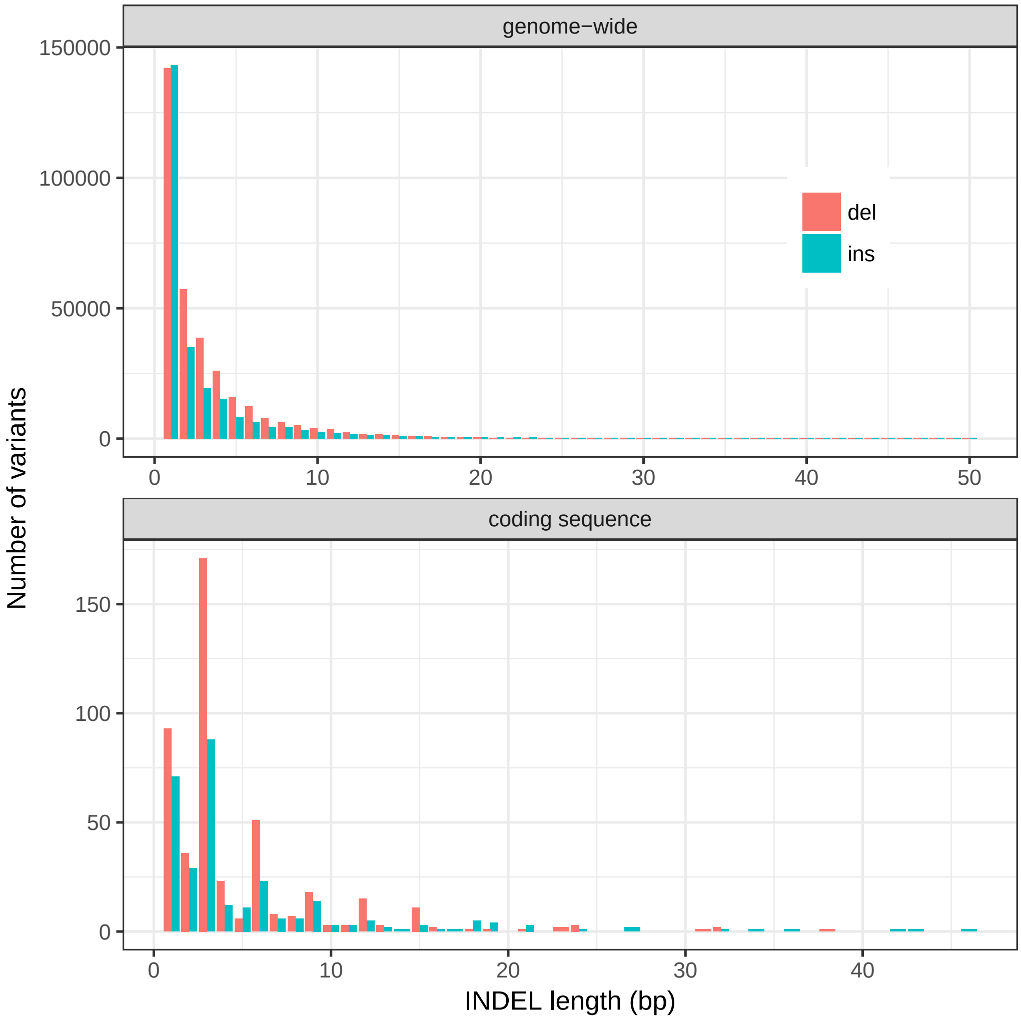


FIGURE B.2: Length distribution of short INDELs (50bp or less) both genome-wide and in coding sequence within the great tit genome.

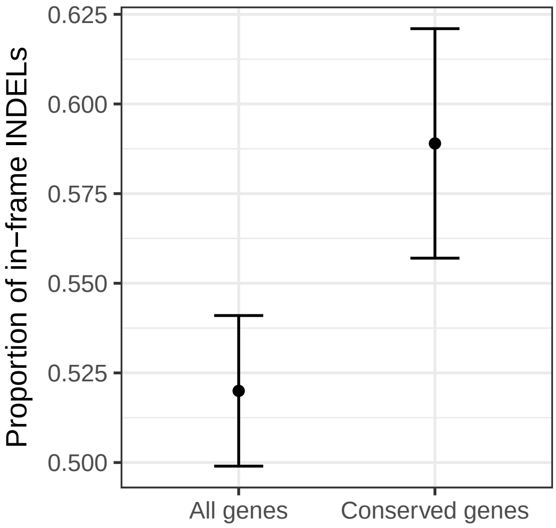


Figure B.3: Proportion of INDELs that preserve the reading frame (INDELs a multiple of 3 in length) in all genes compared to conserved genes.

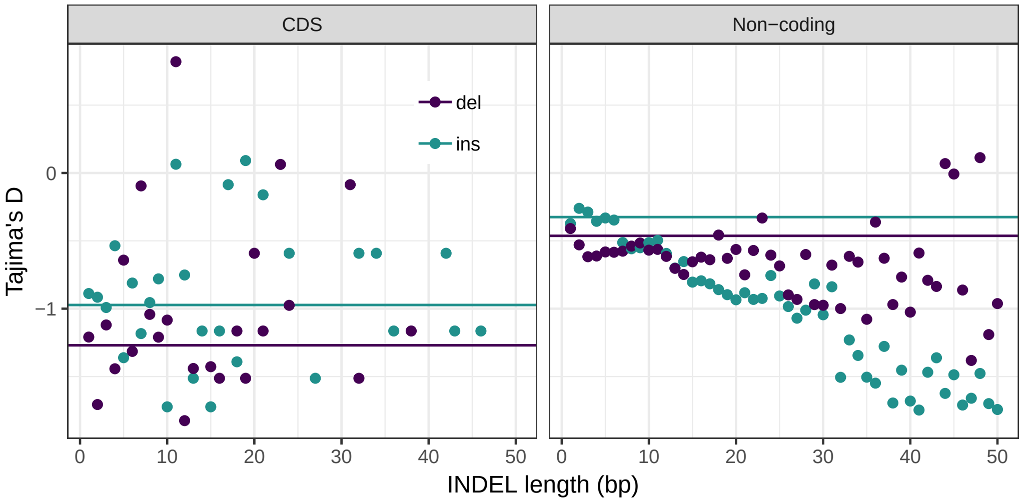


Figure B.4: Tajimas D for insertions and deletions of different lengths in both coding (CDS) and non-coding regions. Horizontal lines represent the estimates when not separated by length.

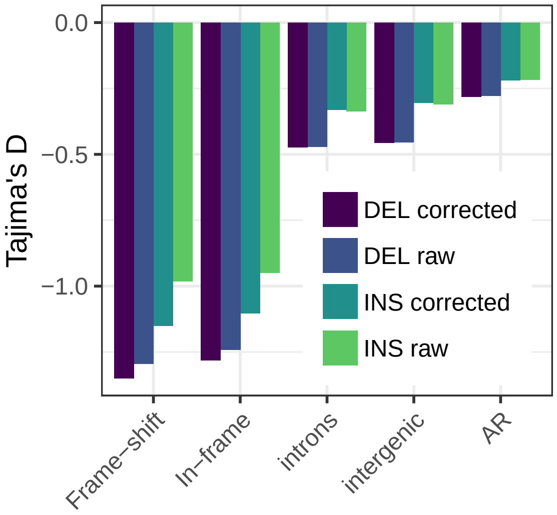


Figure B.5: Tajimas D estimates for insertions and deletions in different genomic contexts, before (raw) and after correcting for polarisation error (corrected).

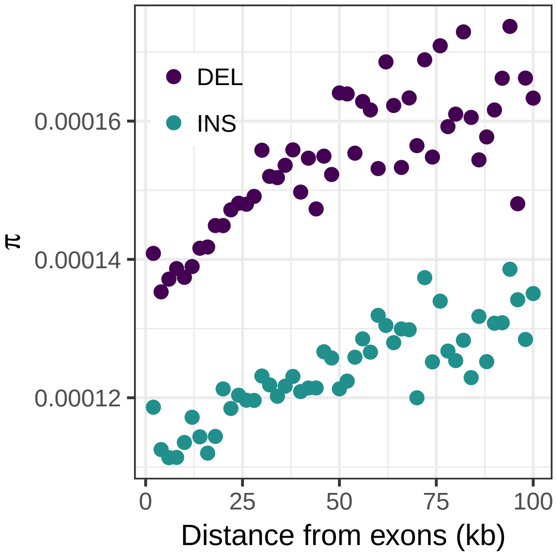


FIGURE B.6: The relationship between INDEL diversity and distance from exons. Each point represents a 2kb bin.

TABLE B.2: Maximum likelihood parameter estimates for models assuming equal mutation rates between neutral and focal sites, fitted to coding sequence INDELs with INDELs in ancestral repeats as neutral reference. Where $AIC = AIC_{bestmodel} - AIC_{lowerrankedmode}$.

Model	Variants	C	θ	γ	scale	shape	ϵ	α	ΔAIC
Discrete	insertions	1	4.92×10^{-6}	-1.14	-	-	0.0799		
$C = 2$	insertions	2	0.000134	-801	-	-	0.000307	71%	
	deletions	1	8.32×10^{-6}	-2.70	-	-	0.0368		
	deletions	2	0.000206	-649	-	-	3.12×10^{-7}	85%	0
Discrete	insertions	1	7.32×10^{-7}	10000	-	-	0.00144		
$C = 3$	insertions	2	0.000141	-748	-	-	0.0870		
	insertions	3	4.04×10^{-6}	-2.56	-	-	4.81×10^{-5}	91%	
	deletions	1	1.17×10^{-6}	83.3	-	-	0.00137		
	deletions	2	9.70×10^{-6}	-5.94	-	-	0.0297		
	deletions	3	0.000208	-859	-	-	0.000368	99%	-5.17
Continuous	insertions	1	0.000130	-1498	2284	0.656	0.0685	100%	
	deletions	1	0.000204	-1551	2326	0.667	0.0442	100%	-106
Discrete	insertions	1	0.000103	-37.4	-	-	0.149	100%	
$C = 1$	deletions	1	0.000171	-70.4	-	-	0.0592	100%	-2741

TABLE B.3: Maximum likelihood parameter estimates for models assuming equal mutation rates between neutral and focal sites, fitted to coding sequence INDELs with INDELs in non-coding regions as neutral reference. Where $AIC = AIC_{bestmodel} - AIC_{lowerrankedmode}$.

Model	Variants	C	θ	γ	scale	shape	ϵ	α	ΔAIC
Discrete	insertions	1	4.79×10^{-6}	-0.264	-	-	0.0729		
$C = 2$	insertions	2	0.000156	-897	-	-	0.000526	63%	
	deletions	1	7.79×10^{-6}	-1.70	-	-	0.0366		
	deletions	2	0.000205	-629	-	-	0.00587	79%	0
Discrete	insertions	1	6.59×10^{-6}	-2.23	-	-	0.0202		
$C = 3$	insertions	2	7.91×10^{-5}	-1011	-	-	0.0491		
	insertions	3	7.76×10^{-5}	-2738	-	-	0.453	85%	
	deletions	1	4.90×10^{-6}	0.597	-	-	0.0839		
	deletions	2	1.52×10^{-5}	-21.1	-	-	0.000163		
	deletions	3	0.000195	-2483	-	-	1.43×10^{-5}	100%	-13.8
Continuous	insertions	1	0.000154	-1614	2413	0.669	0.0667	100%	
	deletions	1	0.000209	-1553	2401	0.647	0.0470	100%	-147
Discrete	insertions	1	0.000153	-54.9	-	-	0.141	100%	
$C = 1$	deletions	1	0.000208	-75.7	-	-	0.0648	100%	-3210

TABLE B.4: Maximum likelihood parameter estimates for models with mutation rates free to vary between neutral and focal sites, fitted to non-coding INDELs. Where $AIC = AIC_{bestmodel} - AIC_{lowerrankedmode}$.

Model	Variants	C	θ	γ	scale	shape	ϵ	ΔAIC
Continuous	insertions	1	0.000170	-53.6	1553	0.0345	0.0110	
	deletions	1	0.000293	-75.5	715	0.106	0.0166	0
Discrete	insertions	1	3.32×10^{-5}	4.64	-	-	0.0358	
$C = 2$	insertions	2	0.000122	-1.44	-	-	0.00293	
	deletions	1	0.000113	1.15	-	-	0.00306	
	deletions	2	0.000117	-5.33	-	-	0.0231	-3.43
Discrete	insertions	1	1.20×10^{-5}	99852	-	-	0.463	
$C = 3$	insertions	2	0.000141	-0.514	-	-	0.0241	
	insertions	3	0.000282	-879	-	-	8.59×10^{-6}	
	deletions	1	3.26×10^{-5}	4.67	-	-	0.127	
	deletions	2	0.000169	-127	-	-	0.0327	
	deletions	3	0.000172	-1.65	-	-	0.0294	-10.9
Discrete	insertions	1	0.000161	-0.204	-	-	0.0584	
$C = 1$	deletions	1	0.000223	-0.831	-	-	0.0451	-131

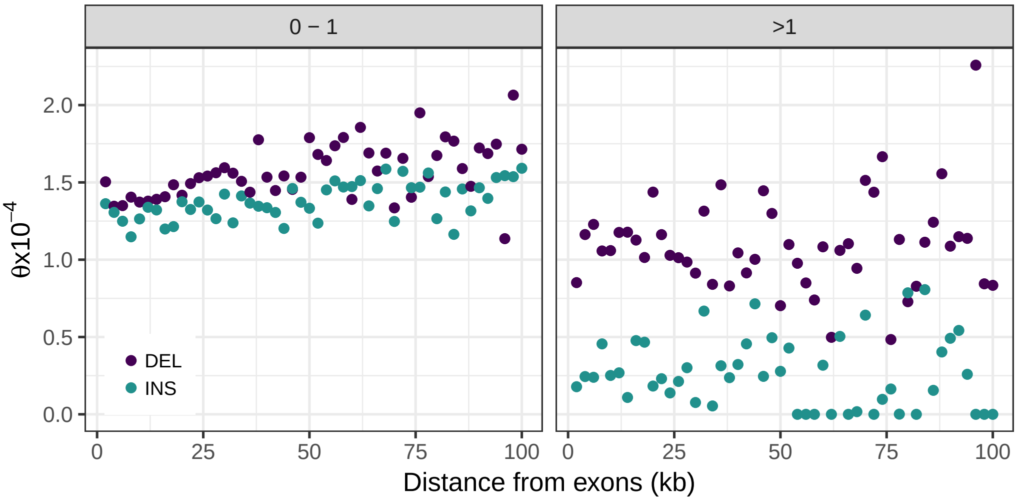


FIGURE B.7: The relationship between and distance from exons for putatively neutral INDELs (left panel, $-\gamma$ between 0 and 1) and negatively selected INDELs (right panel, $-\gamma > 1$). Insertions (INS) in turquoise and deletions (DEL) in purple. Each point represents a 2kb bin.

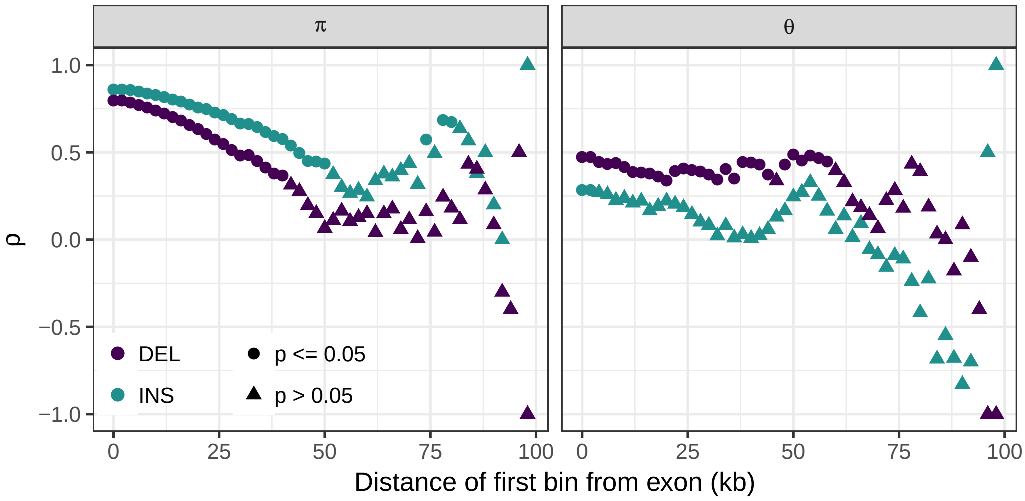


Figure B.8: The strength and significance of Spearman's correlations between distance from exons and nucleotide diversity (left) and model based estimates of the scaled mutation rate (right), with increasingly cumulatively down sampled datasets, reduced by iteratively removing the bin nearest to exons for each correlation.

Table B.5: Maximum likelihood parameter estimates for the best-fit model with mutation rates free to vary between neutral and focal sites, fitted to coding INDELs.

Model	Variants	C	θ	γ	scale	shape	ϵ
Continuous	insertions	1	2.22×10^{-5}	-336	986	0.341	0.0339
	deletions	1	5.38×10^{-5}	-374	758	0.494	0.0169

Appendix C

Supplementary Material to Chapter 4

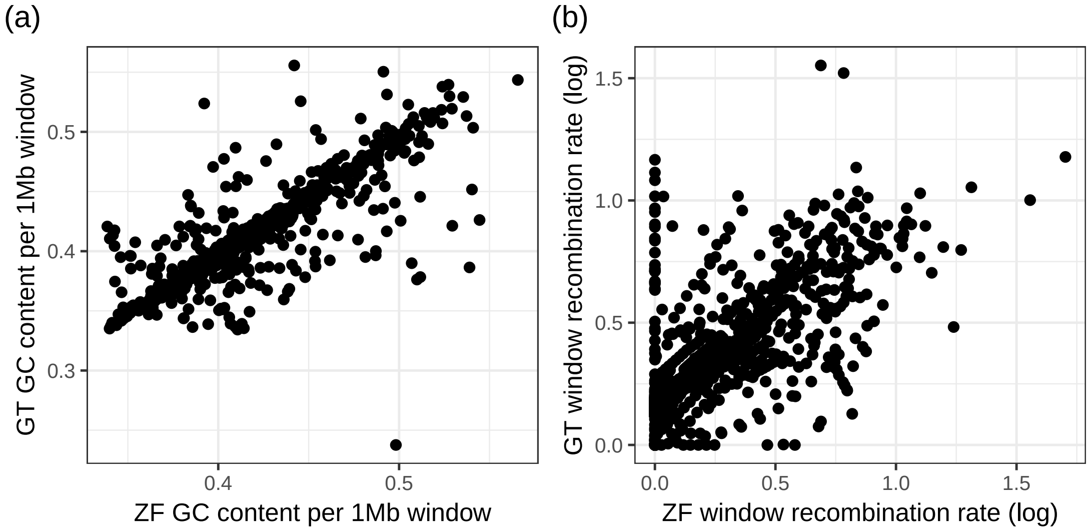


Figure C.1: Relationships between zebra finch GC content and great tit GC content (a) and zebra finch recombination rate and great tit recombination rate (b) across the 1Mb window dataset.

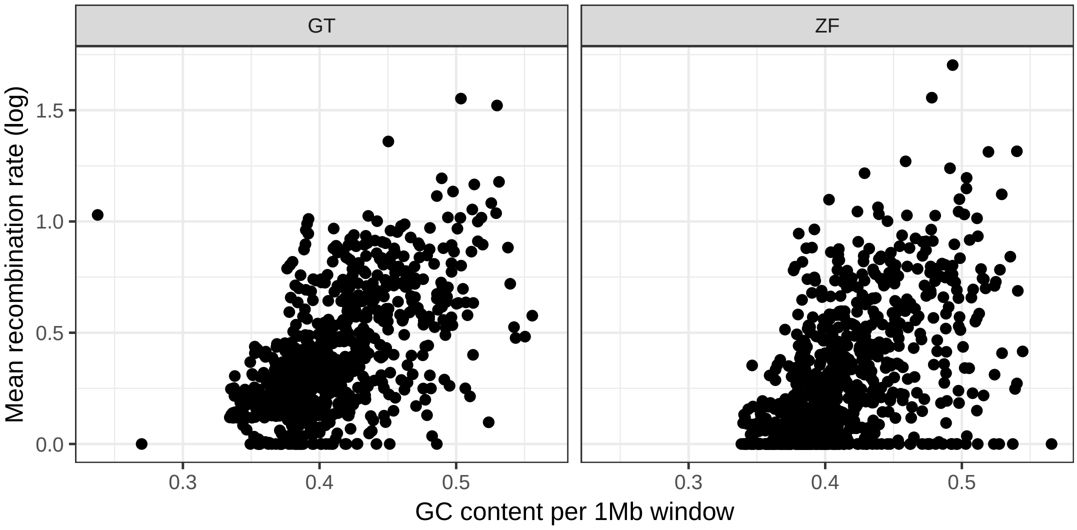


Figure C.2: Relationships between GC content and recombination rate across the 1Mb window dataset in both the great tit (GT) and the zebra finch (ZF).

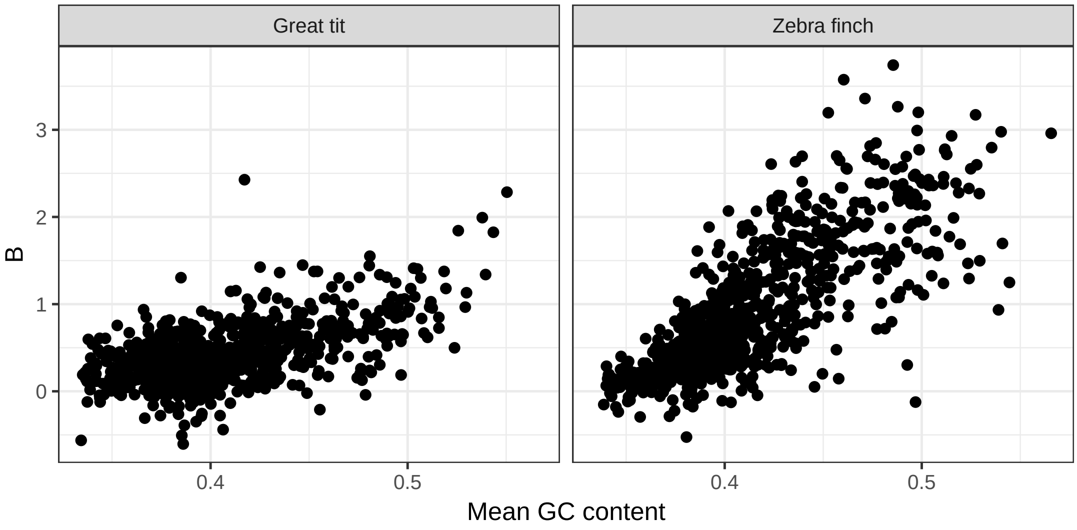


Figure C.3: The relationship between mean window GC content and the strength of gene conversion (B) in the great tit and zebra finch.

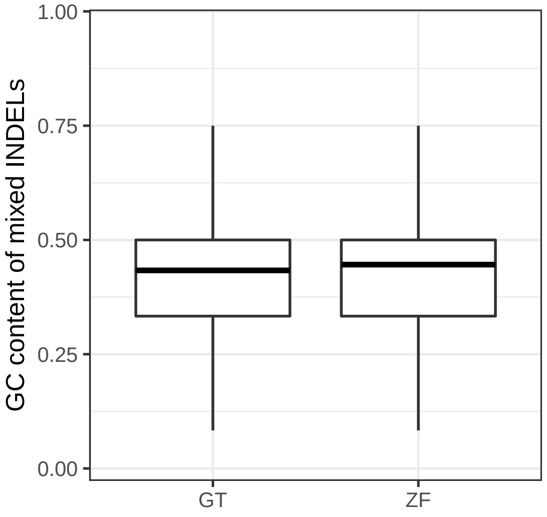


Figure C.4: Distribution of GC content within non-coding mixed INDELs (INDELs comprised of GC and AT bases) for both great tit (GT) and zebra finch (ZF).

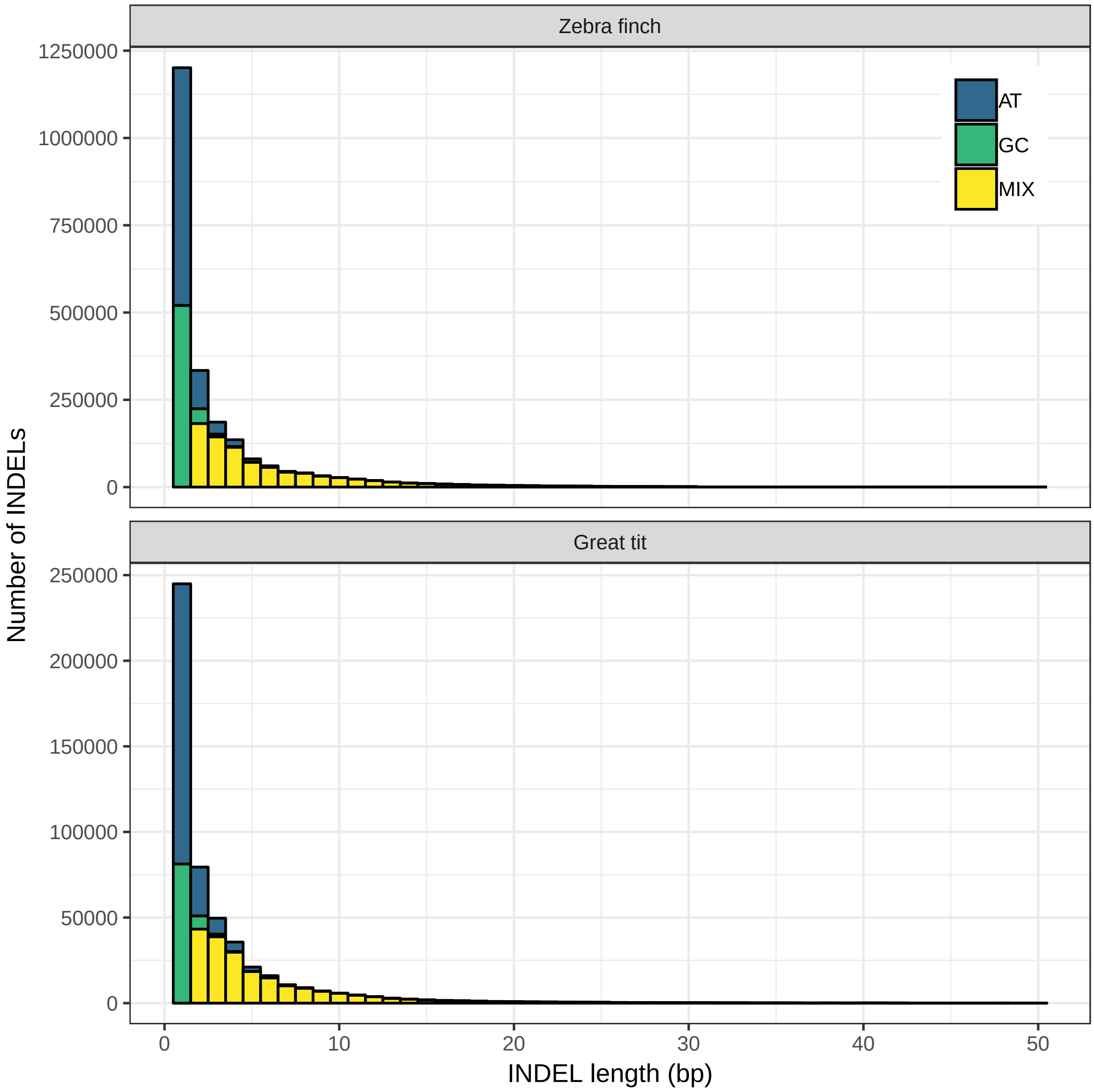


Figure C.5: Length distribution for non-coding INDELs, split by base category (AT, GC or MIX) for both zebra finch and great tit.

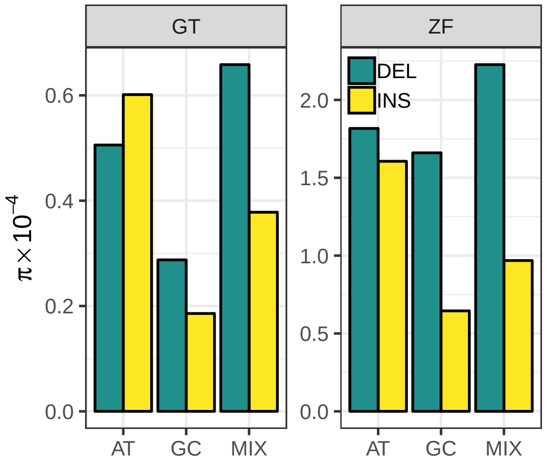


FIGURE C.6: Estimated INDEL diversity (π) for AT, GC and AT and GC mixed (MIX) insertions and deletions in both the great tit (GT) and the zebra finch (ZF).

Table C.1: Maximum likelihood estimates of the population scaled mutation rate ($\theta = 4N_e\mu$), selection coefficient ($\gamma = 4N_e s$) and polarisation error rate (ϵ) for AT, GC and mixed (MIX) content insertions and deletions.

base content	species	INDEL type	θ	γ	ϵ
AT	great tit	insertions	8.24×10^{-5}	-0.391	0
AT	great tit	deletions	6.89×10^{-5}	-0.141	0.0165
GC	great tit	insertions	2.32×10^{-5}	0.353	0.0148
GC	great tit	deletions	4.25×10^{-5}	-0.717	0
MIX	great tit	insertions	4.83×10^{-5}	0.109	4.80×10^{-6}
MIX	great tit	deletions	0.000106	-1.15	0.00226
AT	zebra finch	insertions	0.000504	0.313	0.00261
AT	zebra finch	deletions	0.000554	0.717	0.00273
GC	zebra finch	insertions	0.000167	3.27	0
GC	zebra finch	deletions	0.000541	-0.0927	0.00178
MIX	zebra finch	insertions	0.000260	> 10	0.0230
MIX	zebra finch	deletions	0.000843	-1.22	3.56×10^{-14}

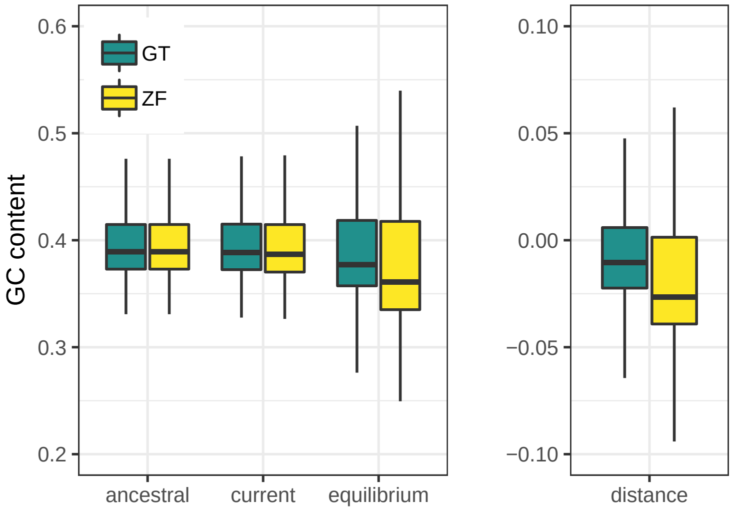


Figure C.7: Per window estimates of ancestral GC content, current GC content, equilibrium GC content and the present distance from equilibrium for both species.

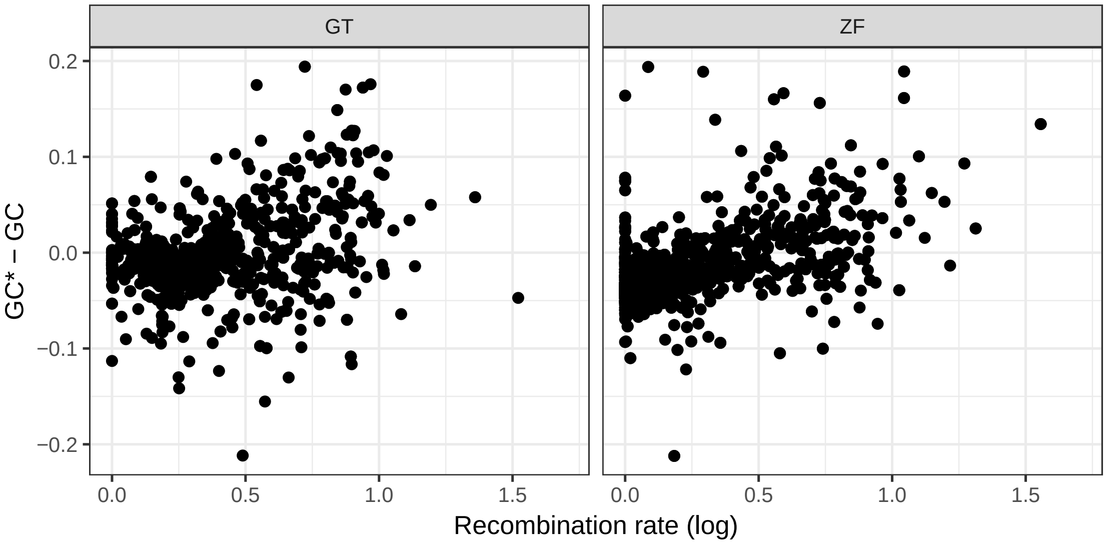


Figure C.8: The relationship between distance from equilibrium GC (GC∗ - GC) and recombination rate in both the great tit (GT) and the zebra finch (ZF).

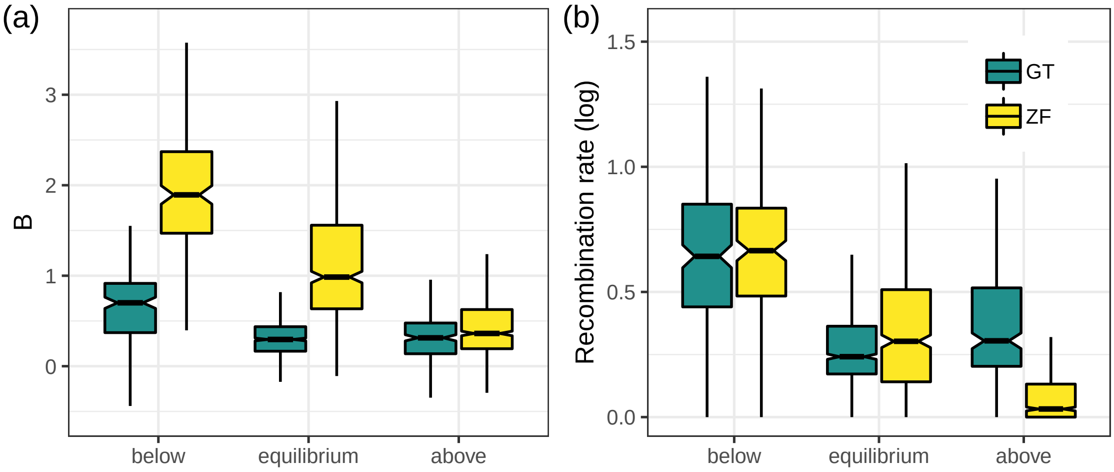


FIGURE C.9: Estimates of strength of gene conversion, B (a) and recombination rate (b) in 1Mb windows in the bins of distance from GC equilibrium, 'below' (current GC is below equilibrium), 'equilibrium' (current GC is within 2.5% of equilibrium GC) and 'above' (current GC is above equilibrium GC) for both the great tit (GT) and zebra finch (ZF)

Table C.2: Correlations between INDEL ΔGC and other genomic variables across the window dataset.

variable	species	INDEL type	correlation	p value	method
recombination	great tit	INS	0.40	2.95×10^{-30}	spearman
recombination	zebra finch	INS	0.35	3.48×10^{-23}	spearman
recombination	great tit	DEL	-0.61	3.64×10^{-78}	spearman
recombination	zebra finch	DEL	-0.58	2.89×10^{-69}	spearman
recombination	great tit	INDEL	-0.085	0.0220	spearman
recombination	zebra finch	INDEL	-0.49	7.48×10^{-46}	spearman
B	great tit	INS	0.33	7.59×10^{-21}	spearman
B	zebra finch	INS	0.54	$< 2.2 \times 10^{-16}$	spearman
B	great tit	DEL	-0.49	$< 2.2 \times 10^{-16}$	spearman
B	zebra finch	DEL	-0.82	$< 2.2 \times 10^{-16}$	spearman
B	great tit	INDEL	-0.054	0.146	spearman
B	zebra finch	INDEL	-0.62	$< 2.2 \times 10^{-16}$	spearman

Table C.3: GC content in coding and non-coding regions in the great tit and zebra finch genomes.

region	species	GC
non-coding	great tit	0.41
coding	great tit	0.52
non-coding	zebra finch	0.41
coding	zebra finch	0.48

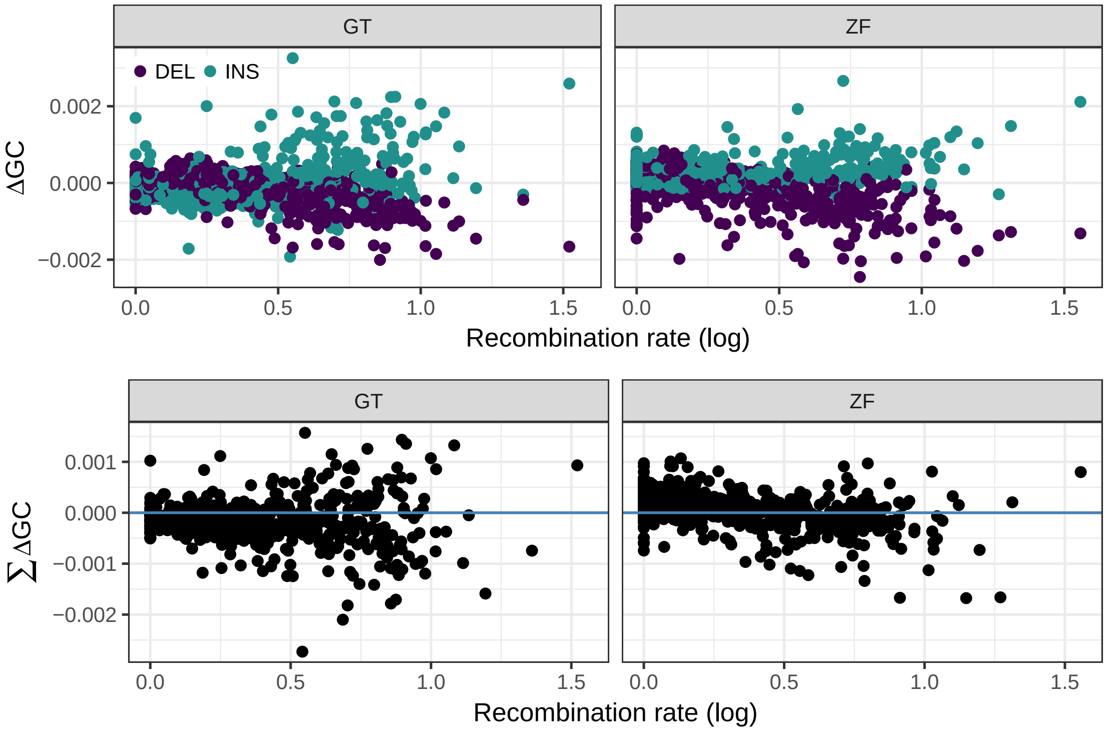


FIGURE C.10: The relationship of ΔGC for insertions and deletions (top row) and delta GC for combined INDELs ($\sum \Delta GC$, bottom row) against recombination rate in both great tit (GT) and zebra finch (ZF).

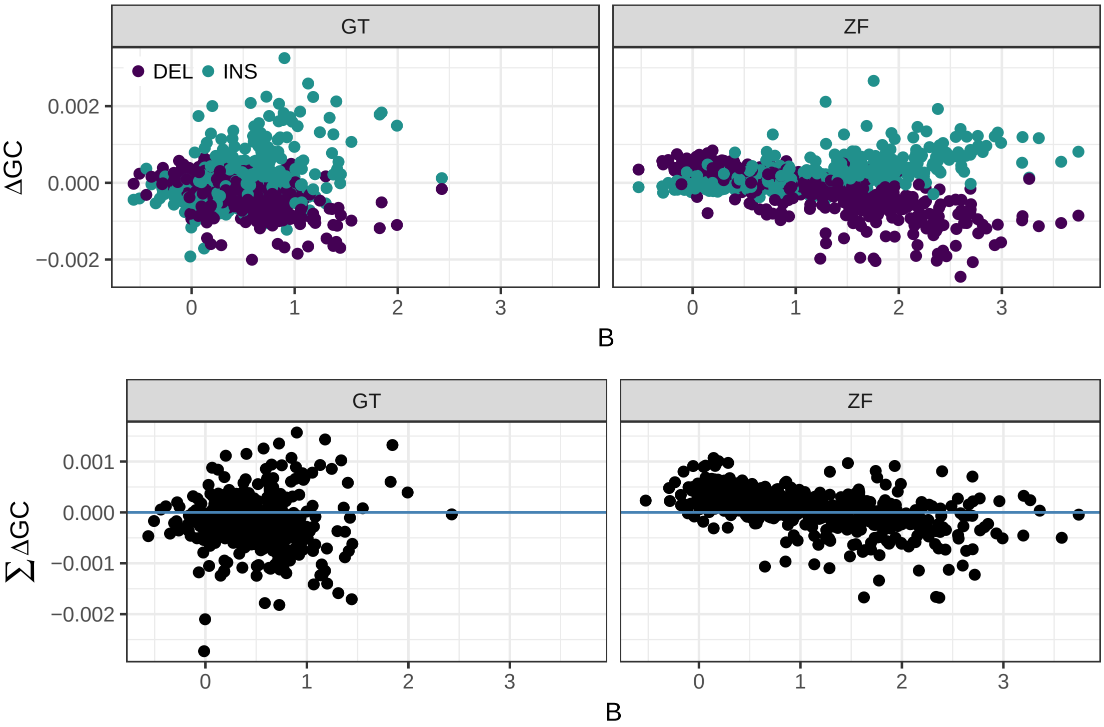


Figure C.11: The relationship of ΔGC for insertions and deletions (top row) and delta GC for combined INDELs ($\sum \Delta GC$, bottom row) against the strength of gBGC (B) in both great tit (GT) and zebra finch (ZF).

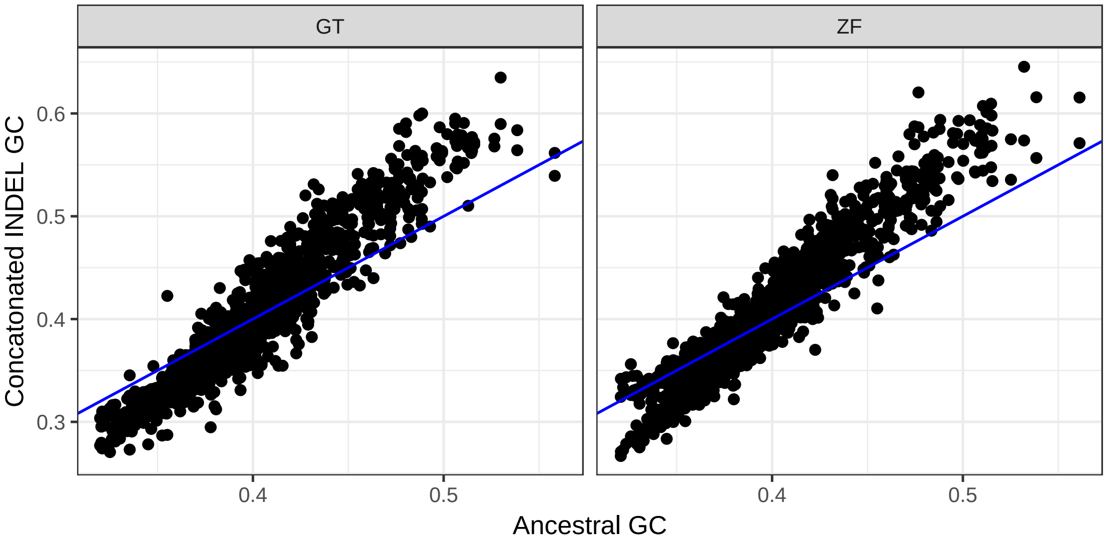


Figure C.12: The relationship between ancestral GC content and INDEL GC content (from concatenated fixed INDELs) per window in both great tit (GT) and zebra finch (ZF) from the alignment dataset. The blue line is the y=x line.

Bibliography

1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319): 1061–1073.

Ananda, G., Walsh, E., Jacob, K. D., Krasilnikova, M., Eckert, K. A., Chiaromonte, F., and Makova, K. D. 2013. Distinct Mutational Behaviors Differentiate Short Tandem Repeats from Microsatellites in the Human Genome. *Genome Biology and Evolution*, 5(3): 606–620.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062): 1149–52.

Andolfatto, P., Wong, K. M., and Bachtrog, D. 2011. Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species. *Genome Biol Evol*, 3: 114–28.

Arbeithuber, B., Betancourt, A. J., Ebner, T., and Tiemann-Boege, I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences*, 112(7): 2109–2114.

Axelsson, E., Webster, M. T., Smith, N. G., Burt, D. W., and Ellegren, H. 2005. Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Research*, 15(1): 120–125.

Backström, N., Forstmeier, W., Schielzeth, H., Mellenius, H., Nam, K., Bolund, E., Webster, M. T., Ost, T., Schneider, M., Kempenaers, B., and Ellegren, H. 2010. The recombination landscape of the zebra finch Taeniopygia guttata genome. *Genome Res*, 20(4): 485–95.

Balakrishnan, C. N. and Edwards, S. V. 2008. Nucleotide Variation, Linkage Disequilibrium and Founder-Facilitated Speciation in Wild Populations of the Zebra Finch (Taeniopygia guttata). *Genetics*, 181(2): 645–660.

Barker, F. K., Cibois, A., Schikler, P., Feinstein, J., and Cracraft, J. 2004. Phylogeny and diversification of the largest avian radiation. *Proceedings of the National Academy of Sciences*, 101(30): 11040–11045.

Barton, H. J. and Zeng, K. 2018. New Methods for Inferring the Distribution of Fitness Effects for INDELs and SNPs. *Molecular Biology and Evolution*, 35(6): 1536–1546.

Barton, H. J. and Zeng, K. 2019. The Impact of Natural Selection on Short Insertion and Deletion Variation in the Great Tit Genome. *Genome Biology and Evolution*, 11(6): 1514–1524.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and Massy, B. d. 2010. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science*, 327(5967): 836–840.

Benzer, S. 1961. On the topography of the genetic fine structure. *Proceedings of the National Academy of Sciences*, 47(3): 403–415.

Besenbacher, S., Liu, S., Izarzugaza, J. M. G., Grove, J., Belling, K., Bork-Jensen, J., Huang, S., Als, T. D., Li, S., Yadav, R., Rubio-García, A., Lescai, F., Demontis, D., Rao, J., Ye, W., Mailund, T., Friborg, R. M., Pedersen, C. N. S., Xu, R., Sun, J., Liu, H., Wang, O., Cheng, X., Flores, D., Rydza, E., Rapacki, K., Damm Sørensen, J., Chmura, P., Westergaard, D., Dworzynski, P., Sørensen, T. I. A., Lund, O., Hansen, T., Xu, X., Li, N., Bolund, L., Pedersen, O., Eiberg, H., Krogh, A., Børglum, A. D., Brunak, S., Kristiansen, K., Schierup, M. H., Wang, J., Gupta, R., Villesen, P., and Rasmussen, S. 2015. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun*, 6: 5969.

Bierne, N. and Eyre-Walker, A. 2006. Variation in synonymous codon use and DNA polymorphism within the Drosophila genome. *Journal of Evolutionary Biology*, 19(1): 1–11.

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research*, 8(7): 1499–1504.

Blake, R. D., Hess, S. T., and Nicholson-Tuell, J. 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *Journal of molecular evolution*, 34(3): 189–200.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., and Miller, W. 2004. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research*, 14(4): 708–715.

Bolívar, P., Mugal, C. F., Nater, A., and Ellegren, H. 2016. Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill–Robertson interference, in an avian system. *Molecular biology and evolution*, 33(1): 216–227.

Bolívar, P., Mugal, C. F., Rossi, M., Nater, A., Wang, M., Dutoit, L., and Ellegren, H. 2018. Biased Inference of Selection Due to GC-Biased Gene Conversion and the Rate of Protein Evolution in Flycatchers When Accounting for It. *Molecular Biology and Evolution*, 35(10): 2475–2486.

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., and Mugal, C. F. 2019. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology*, 20(1): 5.

Borges, R., Szöllősi, G., and Kosiol, C. 2018. Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. *bioRxiv*, page 380246.

Boschiero, C., Gheyas, A. A., Ralph, H. K., Eory, L., Paton, B., Kuo, R., Fulton, J., Preisinger, R., Kaiser, P., and Burt, D. W. 2015. Detection and characterization of small insertion and deletion genetic variants in modern layer chicken genomes. *BMC Genomics*, 16(1).

Brandstrom, M. and Ellegren, H. 2007. The Genomic Landscape of Short Insertion and Deletion Polymorphisms in the Chicken (Gallus gallus) Genome: A High Frequency of Deletions in Tandem Duplicates. *Genetics*, 176(3): 1691–1701.

Britten, R. J. 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proceedings of the National Academy of Sciences*, 99(21): 13633–13635.

Britten, R. J., Rowen, L., Williams, J., and Cameron, R. A. 2003. Majority of divergence between closely related DNA samples is due to indels. *Proceedings of the National Academy of Sciences*, 100(8): 4661–4665.

Burt, D. W. 2002. Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research*, 96(1-4): 97–112.

Bustamante, C. D., Wakeley, J., Sawyer, S., and Hartl, D. L. 2001. Directional selection and the site-frequency spectrum. *Genetics*, 159(4): 1779–88.

C. elegans Sequencing Consortium 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science (New York, N.Y.)*, 282(5396): 2012–2018.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. 2007. Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10): 762–775.

Chintalapati, M., Dannemann, M., and Prüfer, K. 2017. Using the Neandertal genome to study the evolution of small insertions and deletions in modern humans. *BMC Evolutionary Biology*, 17.

Chu, G. 1997. Double strand break repair. *Journal of Biological Chemistry*, 272(39): 24097–24100.

Comeron, J. M. and Kreitman, M. 2000. The Correlation Between Intron Length and Recombination in Drosophila: Dynamic Equilibrium Between Mutational and Selective Forces. *Genetics*, 156(3): 1175–1190.

Corcoran, P., Gossmann, T. I., Barton, H. J., Great Tit HapMap Consortium, Slate, J., and Zeng, K. 2017. Determinants of the Efficacy of Natural Selection on Coding and Noncoding Variability in Two Passerine Species. *Genome Biol Evol*, 9(11): 2987–3007.

Cutter, A. D. and Payseur, B. A. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, 14(4): 262–274.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5): 491–498.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, 12(6): 640–649.

Duret, L. and Arndt, P. F. 2008. The Impact of Recombination on Nucleotide Substitutions in the Human Genome. *PLOS Genetics*, 4(5): e1000071.

Duret, L. and Galtier, N. 2009. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics*, 10(1): 285–311.

Duret, L. and Mouchiroud, D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proceedings of the National Academy of Sciences*, 96(8): 4482–4487.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics*, 162(4): 1837–1847.

Earl, D., Nguyen, N., Hickey, G., Harris, R. S., Fitzgerald, S., Beal, K., Seledtsov, I., Molodtsov, V., Raney, B. J., Clawson, H., Kim, J., Kemena, C., Chang, J.-M., Erb, I., Poliakov, A., Hou, M., Herrero, J., Kent, W. J., Solovyev, V., Darling, A. E., Ma, J., Notredame, C., Brudno, M., Dubchak, I., Haussler, D., and Paten, B. 2014. Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Research*, 24(12): 2077–2089.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., Künstner, A., Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., Uebbing, S., and Wolf, J. B. W. 2012. The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, 491(7426): 756–760.

Eyre-Walker, A. 2002. Changing Effective Population Size and the McDonald-Kreitman Test. *Genetics*, 162(4): 2017–2024.

Eyre-Walker, A. and Hurst, L. D. 2001. The evolution of isochores. *Nature Reviews Genetics*, 2(7): 549.

Eyre-Walker, A. and Keightley, P. D. 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8): 610–618.

Eyre-Walker, A. and Keightley, P. D. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*, 26(9): 2097–108.

Eyre-Walker, A., Woolfit, M., and Phelps, T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173(2): 891–900.

Fryxell, K. J. and Zuckerkandl, E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Molecular Biology and Evolution*, 17(9): 1371–1383.

Galtier, N. 2005. Mutation hot spots in mammalian mitochondrial DNA. *Genome Research*, 16(2): 215–222.

Galtier, N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLOS Genetics*, 12(1): e1005774.

Galtier, N. and Duret, L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23(6): 273–277.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., and Duret, L. 2018. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, 35(5): 1092–1103.

Garcia-Diaz, M. and Kunkel, T. A. 2006. Mechanism of a genetic glissando*: structural biology of indel mutations. *Trends in Biochemical Sciences*, 31(4): 206–214.

Genome 10K Community of Scientists 2009. Genome 10k: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of Heredity*, 100(6): 659–674.

Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. 2015. Quantification of GC-biased gene conversion in the human genome. *Genome Research*, 25(8): 1215–1228.

Gossmann, T. I., Bockwoldt, M., Diringer, L., Schwarz, F., and Schumann, V.-F. 2018. Evidence for Strong Fixation Bias at 4-fold Degenerate Sites Across Genes in the Great Tit Genome. *Frontiers in Ecology and Evolution*, 6.

Gregory, T. 2003. Is small indel bias a determinant of genome size? *Trends in Genetics*, 19(9): 485–488.

Gregory, T. 2004. Insertion–deletion biases and the evolution of genome size. *Gene*, 324: 15–34.

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Houle, D., Charlesworth, B., and Keightley, P. D. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature*, 445(7123): 82–5.

Hansson, B., Ljungqvist, M., Dawson, D. A., Mueller, J. C., Olano-Marin, J., Ellegren, H., and Nilsson, J.-A. 2010. Avian genome evolution: insights from a linkage map of the blue tit (Cyanistes caeruleus). *Heredity*, 104(1): 67–78.

Harris, R. S. 2007. Improved pairwise alignment of genomic DNA. *Ph.D. Thesis, The Pennsylvania State University.*

Hartfield, M. and Keightley, P. D. 2012. Current hypotheses for the evolution of sex and recombination. *Integr Zool*, 7(2): 192–209.

Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol*, 24(8): 1792–800.

Hess, S. T., Blake, J. D., and Blake, R. D. 1994. Wide variations in neighbor-dependent substitution rates. *Journal of Molecular Biology*, 236(4): 1022–1033.

Hodgkinson, A. and Eyre-Walker, A. 2011. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.*, 12(11): 756–766.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F. X., Van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y.-L. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, 43(5): 476–481.

Hu, T. T., Eisen, M. B., Thornton, K. R., and Andolfatto, P. 2013. A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. *Genome Research*, 23(1): 89–98.

Hwang, D. G. and Green, P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39): 13994–14001.

i5K Consortium 2013. The i5k Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *The Journal of Heredity*, 104(5): 595–600.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860–921.

Jackson, B. C., Campos, J. L., and Zeng, K. 2015. The effects of purifying selection on patterns of genetic differentiation between Drosophila melanogaster populations. *Heredity*, 114(2): 163–174.

Jackson, B. C., Campos, J. L., Haddrill, P. R., Charlesworth, B., and Zeng, K. 2017. Variation in the Intensity of Selection on Codon Bias over Time Causes Contrasting Patterns of Base Composition Evolution in Drosophila. *Genome Biology and Evolution*, 9(1): 102–123.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., Fonseca, R. R. d., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M. V., Alfaro-Núñez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jønsson, K. A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T. P., and Zhang, G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215): 1320–1331.

Johnson, K. P. 2003. Deletion Bias in Avian Introns over Evolutionary Timescales. *Molecular Biology and Evolution*, 21(3): 599–602.

Keightley, P. D. and Eyre-Walker, A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177(4): 2251–61.

Keightley, P. D. and Eyre-Walker, A. 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1544): 1187–1193.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. L. 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res*, 19(7): 1195–201.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences*, 100(20): 11484–11489.

Kim, B. Y., Huber, C. D., and Lohmueller, K. E. 2017. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics*, 206(1): 345–361.

Kim, S. 2015. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communications for statistical applications and methods*, 22(6): 665–674.

Klintschar, M. and Wiegand, P. 2003. Polymerase slippage in relation to the uniformity of tetrameric repeat stretches. *Forensic Science International*, 135(2): 163–166.

Kousathanas, A. and Keightley, P. D. 2013. A Comparison of Models to Infer the Distribution of Fitness Effects of New Mutations. *Genetics*, 193(4): 1197–1208.

Kunstner, A., Nabholz, B., and Ellegren, H. 2011. Significant Selective Constraint at 4-Fold Degenerate Sites in the Avian Genome and Its Consequence for Detection of Positive Selection. *Genome Biology and Evolution*, 3(0): 1381–1389.

Kvikstad, E. M. and Duret, L. 2014. Strong heterogeneity in mutation rate causes misleading hallmarks of natural selection on indel mutations in the human genome. *Mol Biol Evol*, 31(1): 23–36.

Laine, V. N., Gossmann, T. I., Schachtschneider, K. M., Garroway, C. J., Madsen, O., Verhoeven, K. J. F., de Jager, V., Megens, H.-J., Warren, W. C., Minx, P., Crooijmans, R. P. M. A., Corcoran, P., Great Tit HapMap Consortium, Sheldon, B. C., Slate, J., Zeng, K., van Oers, K., Visser, M. E., and Groenen, M. A. M. 2016. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat Commun*, 7: 10474.

Leushkin, E. V. and Bazykin, G. A. 2013. Short indels are subject to insertion-biased gene conversion. *Evolution*, 67(9): 2604–13.

Leushkin, E. V., Bazykin, G. A., and Kondrashov, A. S. 2013. Strong mutational bias toward deletions in the Drosophila melanogaster genome is compensated by selection. *Genome Biol Evol*, 5(3): 514–24.

Levinson, G. and Gutman, G. A. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, 4(3): 203–221.

Li, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20): 2843–51.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16): 2078–2079.

Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S. F., Guo, W., Patterson, C., Gregory, C., Strauss, C., Stone, C., Berne, C., Kysela, D., Shoemaker, W. R., Muscarella, M. E., Luo, H., Lennon, J. T., Brun, Y. V., and Lynch, M. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nature Ecology & Evolution*, 2(2): 237–240.

Matsumoto, T., Akashi, H., and Yang, Z. 2015. Evaluation of Ancestral Sequence Reconstruction Methods to Infer Nonstationary Patterns of Nucleotide Substitution. *Genetics*, 200(3): 873–890.

McDonald, J. H. and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328): 652–654.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9): 1297–1303.

McVean, G. a. T. and Charlesworth, B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genetics Research*, 74(2): 145–158.

McVey, M., LaRocque, J. R., Adams, M. D., and Sekelsky, J. J. 2004. Formation of deletions during double-strand break repair in Drosophila DmBlm mutants occurs after strand invasion. *Proceedings of the National Academy of Sciences of the United States of America*, 101(44): 15694–15699.

Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., Ananda, G., Howie, B., Karczewski, K. J., Smith, K. S., Anaya, V., Richardson, R., Davis, J., 1000 Genomes Project Consortium, MacArthur, D. G., Sidow, A., Duret, L., Gerstein, M., Makova, K. D., Marchini, J., McVean, G., and Lunter, G. 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res*, 23(5): 749–61.

Mugal, C. F., Arndt, P. F., and Ellegren, H. 2013. Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. *Molecular Biology and Evolution*, 30(7): 1700–1712.

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., and Glémin, S. 2011. GC-Biased Gene Conversion and Selection Affect GC Content in the Oryza Genus (rice). *Molecular Biology and Evolution*, 28(9): 2695–2706.

Myers, S., Fefferman, C., and Patterson, N. 2008. Can one learn history from the allelic spectrum? *Theor Popul Biol*, 73(3): 342–8.

Nabholz, B., Kunstner, A., Wang, R., Jarvis, E. D., and Ellegren, H. 2011. Dynamic Evolution of Base Composition: Causes and Consequences in Avian Phylogenomics. *Molecular Biology and Evolution*, 28(8): 2197–2210.

Nam, K. and Ellegren, H. 2012. Recombination Drives Vertebrate Genome Contraction. *PLoS Genetics*, 8(5): e1002680.

Ometto, L., Stephan, W., and Lorenzo, D. D. 2005. Insertion/Deletion and Nucleotide Polymorphism Data Reveal Constraints in Drosophila melanogaster Introns and Intergenic Regions. *Genetics*, 169(3): 1521–1527.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., and Lynch, M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science (New York, N.Y.)*, 327(5961): 92–94.

Paradis, E., Claude, J., and Strimmer, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2): 289–290.

Parsch, J. 2003. Selective Constraints on Intron Evolution in Drosophila. *Genetics*, 165(4): 1843–1851.

Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M., and 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(6): 1226–34.

Parvanov, E. D., Petkov, P. M., and Paigen, K. 2010. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science*, 327(5967): 835–835.

Paśko, Ł., Ericson, P. G., and Elzanowski, A. 2011. Phylogenetic utility and evolution of indels: A study in neognathous birds. *Molecular Phylogenetics and Evolution*, 61(3): 760–771.

Petrov, D. A. 2002a. DNA loss and evolution of genome size in Drosophila. *Genetica*, 115(1): 81–91.

Petrov, D. A. 2002b. Mutational Equilibrium Model of Genome Size Evolution. *Theoretical Population Biology*, 61(4): 531–544.

Petrov, D. A. and Hartl, D. L. 1998. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. *Mol Biol Evol*, 15(3): 293–302.

Pool, J. E., Corbett-Detig, R. B., Sugino, R. P., Stevens, K. A., Cardeno, C. M., Crepeau, M. W., Duchen, P., Emerson, J. J., Saelao, P., Begun, D. J., and Langley, C. H. 2012. Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture. *PLoS Genet*, 8(12): e1003080.

Presgraves, D. C. 2006. Intron Length Evolution in Drosophila. *Molecular Biology and Evolution*, 23(11): 2203–2213.

Primmer, C. R., Raudsepp, T., Chowdhary, B. P., Møller, A. P., and Ellegren, H. 1997. Low frequency of microsatellites in the avian genome. *Genome Research*, 7(5): 471–482.

Ptak, S. E. and Petrov, D. A. 2002. How intron splicing affects the deletion and insertion profile in Drosophila melanogaster. *Genetics*, 162(3): 1233–44.

R Core Team 2015. R: A Language and Environment for Statistical Computing.

Rajic, Z. A., Jankovic, G. M., Vidovic, A., Milic, N. M., Skoric, D., Pavlovic, M., and Lazarevic, V. 2005. Size of the protein-coding genome and rate of molecular evolution. *Journal of Human Genetics*, 50(5): 217–229.

Rao, Y. S., Wang, Z. F., Chai, X. W., Wu, G. Z., Nie, Q. H., and Zhang, X. Q. 2010. Indel segregating within introns in the chicken genome are positively correlated with the recombination rates: Indel segregating within introns in the chicken genome. *Hereditas*, 147(2): 53–57.

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M. T. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552): 2571–2580.

Romiguier, J., Ranwez, V., Douzery, E. J. P., and Galtier, N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Research*, 20(8): 1001–1009.

Rousselle, M., Laverré, A., Figuet, E., Nabholz, B., and Galtier, N. 2019. Influence of Recombination and GC-biased Gene Conversion on the Adaptive and Nonadaptive Substitution Rate in Mammals versus Birds. *Molecular Biology and Evolution*, 36(3): 458–471.

Sanger Communications Team 2018. Genetic code of 66,000 UK species to be sequenced.

Sawyer, S. A. and Hartl, D. L. 1992. Population genetics of polymorphism and divergence. *Genetics*, 132(4): 1161–76.

Schneider, A., Charlesworth, B., Eyre-Walker, A., and Keightley, P. D. 2011. A Method for Inferring the Rate of Occurrence and Fitness Effects of Advantageous Mutations. *Genetics*, 189(4): 1427–1437.

Schrider, D. R., Houle, D., Lynch, M., and Hahn, M. W. 2013. Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. *Genetics*, 194(4): 937–54.

Ségurel, L., Wyman, M. J., and Przeworski, M. 2014. Determinants of Mutation Rate Variation in the Human Germline. *Annual Review of Genomics and Human Genetics*, 15(1): 47–70.

Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G., and Peden, J. F. 1995. DNA sequence evolution: the sounds of silence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 349(1329): 241–247.

Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., Hooper, D. M., Strand, A. I., Li, Q., Raney, B., Balakrishnan, C. N., Griffith, S. C., McVean, G., and Przeworski, M. 2015. Stable recombination hotspots in birds. *Science*, 350(6263): 928–32.

Sjödin, P., Bataillon, T., and Schierup, M. H. 2010. Insertion and deletion processes in recent human history. *PLoS One*, 5(1): e8650.

Smeds, L., Mugal, C. F., Qvarnström, A., and Ellegren, H. 2016. High-Resolution Mapping of Crossover and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. *PLOS Genetics*, 12(5): e1006044.

Smit, A. F. A., Hubley, R., and Green, P. 2013. RepeatMasker Open-4.0.

Stapley, J., Birkhead, T. R., Burke, T., and Slate, J. 2008. A Linkage Map of the Zebra Finch Taeniopygia guttata Provides New Insights Into Avian Genome Evolution. *Genetics*, 179(1): 651–667.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., and Smadja, C. M. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Phil. Trans. R. Soc. B*, 372(1736): 20160455.

Sun, C., López Arriaza, J. R., and Mueller, R. L. 2012. Slow DNA Loss in the Gigantic Genomes of Salamanders. *Genome Biology and Evolution*, 4(12): 1340–1348.

Sundström, H., Webster, M. T., and Ellegren, H. 2003. Is the rate of insertion and deletion mutation male biased?: Molecular evolutionary analysis of avian and primate sex chromosome sequences. *Genetics*, 164(1): 259–268.

Syvänen, A.-C. 2005. Toward genome-wide SNP genotyping. *Nature Genetics*, 37 Suppl: S5–10.

Tajima, F. 1983. Evolutionary Relationship of Dna Sequences in Finite Populations. *Genetics*, 105(2): 437–460.

Tajima, F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123(3): 585–595.

Tataru, P., Mollion, M., Glémin, S., and Bataillon, T. 2017. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. *Genetics*, 207(3): 1103–1119.

Taylor, M. S., Ponting, C. P., and Copley, R. R. 2004. Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes. *Genome Research*, 14(4): 555–566.

Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J., and Chen, J.-Q. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, 455(7209): 105–108.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, 43: 11.10.1–33.

van Oers, K., Santure, A. W., De Cauwer, I., van Bers, N. E., Crooijmans, R. P., Sheldon, B. C., Visser, M. E., Slate, J., and Groenen, M. A. 2014. Replicated high-density genetic maps of two great tit populations reveal fine-scale genomic departures from sex-equal recombination rates. *Heredity*, 112(3): 307–316.

Wallberg, A., Glémin, S., and Webster, M. T. 2015. Extreme Recombination Frequencies Shape Genome Variation and Evolution in the Honeybee, Apis mellifera. *PLOS Genetics*, 11(4): e1005189.

Wang, D. and Yu, J. 2011. Both Size and GC-Content of Minimal Introns Are Selected in Human Populations. *PLOS ONE*, 6(3): e17945.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., Searle, S., White, S., Vilella, A. J., Fairley, S., Heger, A., Kong, L., Ponting, C. P., Jarvis, E. D., Mello, C. V., Minx, P., Lovell, P., Velho, T. A. F., Ferris, M., Balakrishnan, C. N., Sinha, S., Blatti, C., London, S. E., Li, Y., Lin, Y.-C., George, J., Sweedler, J., Southey, B., Gunaratne, P., Watson, M., Nam, K., Backström, N., Smeds, L., Nabholz, B., Itoh, Y., Whitney, O., Pfenning, A. R., Howard, J., Völker, M., Skinner, B. M., Griffin, D. K., Ye, L., McLaren, W. M., Flicek, P., Quesada, V., Velasco, G., Lopez-Otin, C., Puente, X. S., Olender, T., Lancet, D., Smit, A. F. A., Hubley, R., Konkel, M. K., Walker, J. A., Batzer, M. A., Gu, W., Pollock, D. D., Chen, L., Cheng, Z., Eichler, E. E., Stapley, J., Slate, J., Ekblom, R., Birkhead, T., Burke, T., Burt, D., Scharff, C., Adam, I., Richard, H., Sultan, M., Soldatov, A., Lehrach, H., Edwards, S. V., Yang, S.-P., Li, X., Graves, T., Fulton, L., Nelson, J., Chinwalla, A., Hou, S., Mardis, E. R., and Wilson, R. K. 2010. The genome of a songbird. *Nature*, 464(7289): 757–762.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2): 256–276.

Weber, C. C., Boussau, B., Romiguier, J., Jarvis, E. D., and Ellegren, H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biology*, 15(12): 549.

Webster, M. T. 2006. Strong Regional Biases in Nucleotide Substitution in the Chicken Genome. *Molecular Biology and Evolution*, 23(6): 1203–1216.

Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G., Patterson, N., Myers, S. R., Curran, J. E., Duggirala, R., Blangero, J., Reich, D., and Przeworski, M. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife*, 4: e04637.

Yan, Y., Yi, G., Sun, C., Qu, L., and Yang, N. 2014. Genome-Wide Characterization of Insertion and Deletion Variation in Chicken Using Next Generation Sequencing. *PLoS ONE*, 9(8): e104652.

Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.-Q., Hurst, L. D., and Tian, D. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature*, 523(7561): 463–7.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8): 1586–1591.

Zhang, G. 2015. Genomics: Bird sequencing project takes off. *Nature*, 522(7554): 34–34.

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., Storz, J. F., Antunes, A., Greenwold, M. J., Meredith, R. W., Ödeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H., Wang, Z., Jin, L., Zhang, P., Hu, H., Yang, W., Hu, J., Xiao, J., Yang, Z., Liu, Y., Xie, Q., Yu, H., Lian, J., Wen, P., Zhang, F., Li, H., Zeng, Y., Xiong, Z., Liu, S., Zhou, L., Huang, Z., An, N., Wang, J., Zheng, Q., Xiong, Y., Wang, G., Wang, B., Wang, J., Fan, Y., da Fonseca, R. R., Alfaro-Núñez, A., Schubert, M., Orlando, L., Mourier, T., Howard, J. T., Ganapathy, G., Pfenning, A., Whitney, O., Rivas, M. V., Hara, E., Smith, J., Farré, M., Narayan, J., Slavov, G., Romanov, M. N., Borges, R., Machado, J. P., Khan, I., Springer, M. S., Gatesy, J., Hoffmann, F. G., Opazo, J. C., Håstad, O., Sawyer, R. H., Kim, H., Kim, K.-W., Kim, H. J., Cho, S., Li, N., Huang, Y., Bruford, M. W., Zhan, X., Dixon, A., Bertelsen, M. F., Derryberry, E., Warren, W., Wilson, R. K., Li, S., Ray, D. A., Green, R. E., O'Brien, S. J., Griffin, D., Johnson, W. E., Haussler, D., Ryder, O. A., Willerslev, E., Graves, G. R., Alström, P., Fjeldså, J., Mindell, D. P., Edwards, S. V., Braun, E. L., Rahbek, C., Burt, D. W., Houde, P., Zhang, Y., Yang, H., Wang, J., Avian Genome Consortium, Jarvis, E. D., Gilbert, M. T. P., and Wang, J. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science (New York, N.Y.)*, 346(6215): 1311–1320.

Zhang, J., Chiodini, R., Badr, A., and Zhang, G. 2011. The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, 38(3): 95–109.

Zhao, Z. and Boerwinkle, E. 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome research*, 12(11): 1679–1686.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. 2014. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A*, 111(4): E455–64.