Article:
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). pp. 1514-1524. ISSN 1759-6653

https://doi.org/10.1093/gbe/evz068

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

Henry J. Barton,*,[1] and Kai Zeng [1]

[1]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, United Kingdom

***Corresponding author:** E-mail: hbarton2@sheffield.ac.uk.

Associate Editor: editor

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.

Key words: insertions, deletions, distribution of fitness effects, linked selection, adaptive mutation

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 $\leq$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, 2002; 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 visa versa), which can confound estimates of selection (Kvikstad and Duret, 2014) (see figure 1 in Barton and Zeng (2018)).

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, 2002; 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).

Materials and Methods

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.

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.

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.

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.

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 S1), and solved the system of linear equations:

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

$$\phi_{n-i}^{del,obs} = (1-\epsilon^{del})\phi_{n-i}^{del} + \epsilon^{ins}\phi_i^{ins} \quad (2)$$

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.

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.

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 S4). 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.

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.

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).

Results

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

Table 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 wide	0.00310	0.000356	0.000113 (0.000112)	0.000142 (0.000144)
Ancestral repeats	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}	-	-	-

Table 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.

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

NOTE.—Where γ values are presented for the continuous model these are mean γ estimates and the product of the scale and shape parameters.

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 S1).

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 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 are a multiple of three in length, in other words, mutations that preserve the reading frame (in-frame) (figure S2). This enrichment is even more pronounced in conserved genes (figure S3). 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

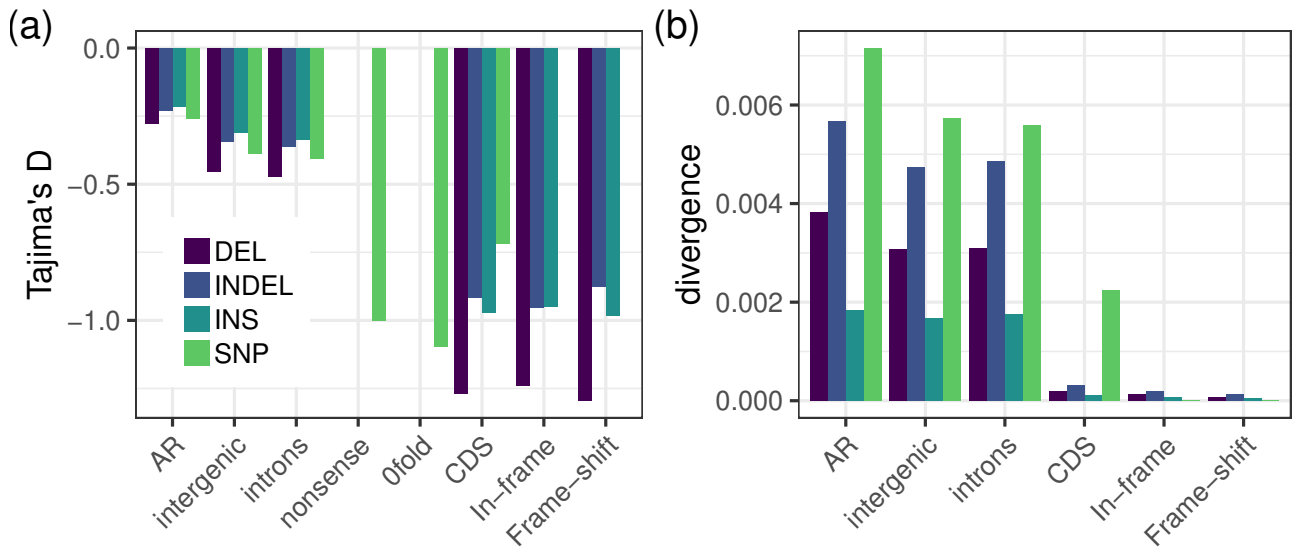

FIG. 1. Tajima's D estimates for SNPs, INDELs (unpolarised), insertions (INS) and deletions (DEL) in different genomic contexts. Divergence estimates for SNPs are presented as the true divergence divided by 10.

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 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 1 and figure 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 1a) is consistent with a recent population expansion for the great tit, as previously reported (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 1a). In non-coding regions, Tajima's D is negatively correlated with INDEL size for both insertions (Spearman's $\rho=-0.95$, $p<2.2\times10^{-16}$) and deletions (Spearman's $\rho=-0.40$, $p=0.0038$), suggesting that longer

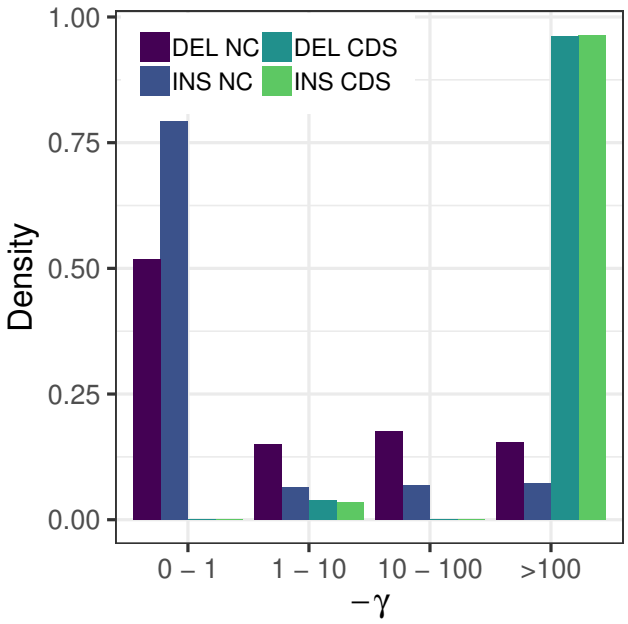

FIG. 2. Distribution of fitness effects for non-coding insertions (INS NC), non-coding deletions (DEL NC), coding insertions (INS CDS) and coding deletions (DEL CDS), shown as the proportion of mutations falling into different selection coefficient (γ) bins.

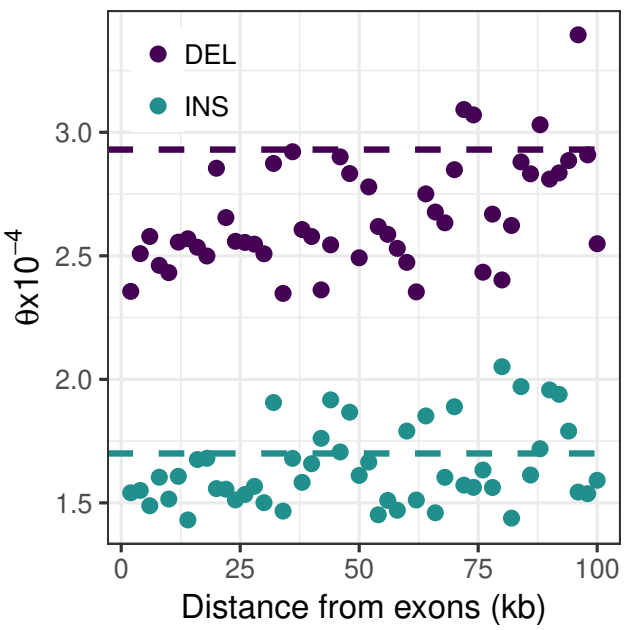

FIG. 3. Relationship between mutation rate estimates (θ) for insertions (turquoise) and deletions (purple) and distance from exons in 2kb windows. Dashed lines represent the genome wide average mutation rate for non-coding variants, as show in table 2.

variants are probably more deleterious (figure S4). In coding regions we lack power when sub-setting INDELs by length (figure S4).

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 1b). These results are robust to polarisation error (table 1, figure S5).

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). The best-fit model for each case is reported in table 2.

The best-fit INDEL DFE (according to AIC, see table S2) 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

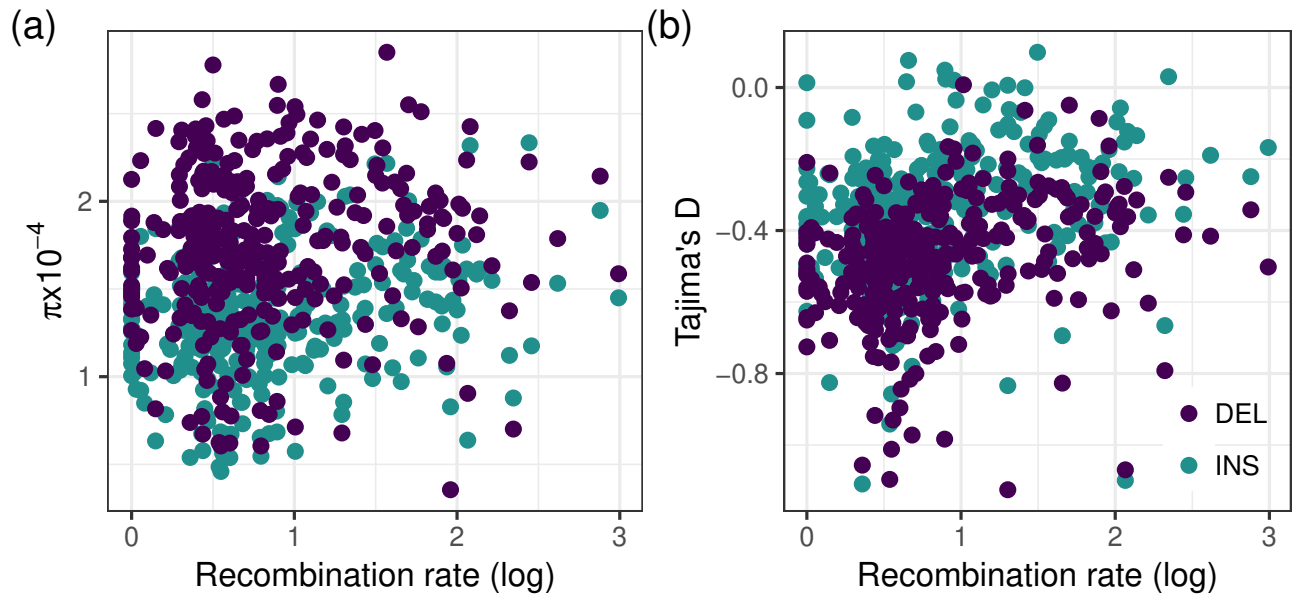

FIG. 4. The relationship between local recombination rate (log transformed) and π (a) and Tajima's D (b) for both insertions (turquoise) and deletions (purple)

deletions for the remaining 4% of sites (figure 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% for insertions and 86% for deletions (table 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 2 and table S3).

The non-coding INDEL data is best fit by a continuous gamma distribution of fitness effects (table S4). We see small shape parameter estimates of 0.0345 for insertions and 0.106 for deletions (table 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 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.

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 ρ=0.28, p=0.046) with distance from exons (figure 3). This relationship is corroborated when using π estimates for deletions and insertions (deletions: Spearman's $\rho=0.79$, $p=2.2\times10^{-16}$, insertions: Spearman's $\rho=0.84$, $p=2.2\times10^{-16}$, see figure S6). 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 $\rho=0.54$, $p=7.9\times10^{-5}$, θ_{ins}: Spearman's $\rho=0.57$, $p=2.3\times10^{-5}$). 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 S7). 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 S8).

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 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 $\rho=0.30$, $p=3.7\times10^{-8}$) and deletions (Spearman's $\rho=0.33$, $p=1.5\times10^{-9}$) (figure 4b).

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.

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 S2). This enrichment of in-frame INDELs is even more pronounced in conserved genes (figure S3). 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 2, figure 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 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, 2002; 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 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 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 S5), 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 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 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 S2, 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.

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 2), corresponding to 76% of insertions and 52% of deletions having γ values between 0 and -1, and thus effectively neutral (figure 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). 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 S8). 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 S7). 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 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 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.

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.

Supplementary Material

Supplementary tables S1-S5 and figures S1-S8 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

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).

References

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.

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.

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.

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.

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.

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.

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.

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.

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 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.

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.

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.

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 Biol Evol*, 9(1): 102–123.

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.

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.

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.

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.

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.

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.

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.

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

Nam, K., Mugal, C., Nabholz, B., Schielzeth, H., Wolf, J. B., Backström, N., Künstner, A., Balakrishnan, C. N., Heger, A., Ponting, C. P., and others 2010. Molecular evolution of genes in avian genomes. *Genome biology*, 11(6): 1.

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.

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.

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

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

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

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.

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.

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

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.

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.

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.

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.

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.

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.

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.