Stable recombination hotspots in birds
Supplementary Online Material

Sonal Singhal*, Ellen M. Leffler*, Keerthi Sannareddy, Isaac Turner, Oliver Venn, Daniel M. Hooper, Alva I. Strand, Qiye Li, Brian Raney, Christopher N. Balakrishnan, Simon C. Griffith, Gil McVean, Molly Przeworski

November 5, 2015

Contents

1 Materials and Methods **3**
1.1 Samples . . . 3
1.2 Reference Genome . . . 3
1.3 Species Tree Inference and TreeMix . . . 4
1.4 Confirming the Absence of PRDM9 . . . 4
1.5 Identifying the Pseudoautosomal Region . . . 5
1.6 Sample Preparation and Sequencing . . . 5
1.7 Read Processing and Alignment . . . 6
1.8 Variant Calling and Filtering . . . 6
1.8.1 Variant Calling . . . 6
1.8.2 Sex Chromosomes . . . 7
1.8.3 Coverage and Repeat Masking Filtering . . . 7
1.8.4 Mendelian Errors . . . 7
1.8.5 Variant Quality . . . 8
1.8.6 Ancestral Allele Inference . . . 8
1.8.7 Estimation of Nucleotide Diversity and the Mutation Rate . . . 9
1.9 Phasing Haplotypes . . . 9
1.9.1 Phasing . . . 9
1.9.2 Phasing Errors . . . 10
1.10 Fine-Scale Recombination Maps . . . 11
1.10.1 Generating Recombination Maps . . . 11
1.10.2 Ancestral Allele for LDhelmet . . . 11
1.10.3 Mutation Matrix Estimation . . . 11
1.10.4 Defining the Block Penalty . . . 12
1.10.5 Comparison to an Existing Genetic Map . . . 13
1.11 Hotspots . . . 13
1.11.1 Power to Detect Hotspots . . . 13
1.11.2 Identifying Hotspots . . . 14
1.11.3 Hotspot Validation . . . 14

1.11.4 False Positive Rate . . . 15
1.11.5 Likelihood of Shared Hotspots . . . 16
1.11.6 Null models for Hotspot Sharing . . . 16
1.11.7 Hotspot conservation across the avian phylogeny . . . 17
1.11.8 Matched hotspots and coldspots . . . 17
1.11.9 Estimation of GC* . . . 18
1.11.10 Motif Discovery . . . 19
1.12 Analysis of Gene Expression . . . 19
1.13 Inversion Discovery . . . 19
1.14 Packages used for Data Visualization and Analysis . . . 20
1.15 Representative Commands . . . 20
1.15.1 Read Mapping . . . 20
1.15.2 Mark Duplicates . . . 20
1.15.3 Re-align Indels . . . 20
1.15.4 Fix Mate-Pair Information . . . 20
1.15.5 Call Raw SNPs . . . 21
1.15.6 Recalibrate Raw SNPs with BQSR . . . 21
1.15.7 Using Cortex for *de novo* Variant Calls . . . 22
1.15.8 Recalibrating Variants with VQSR . . . 22
1.15.9 Identify Mendelian Errors . . . 23
1.15.10 Get Species Genomes . . . 23
1.15.11 Family Phasing . . . 23
1.15.12 Population Phasing . . . 24
1.15.13 Infer Recombination Maps . . . 24
1.15.14 Motif Discovery for Hotspots . . . 24
1.15.15 Identify Inversions . . . 24

2 Supplementary Figures **25**

3 Supplementary Tables **56**

1 Materials and Methods

1.1 Samples

For zebra finch *Taeniopygia guttata* (family: *Estrildidae*), we sequenced twenty adults captured at four localities from the population studied at Fowlers Gap, New South Wales in southeastern Australia (42). Further details on the samples' sex and geographic origins can be found in Table S1. Initial analysis of sequencing results showed that one sample had been duplicated, reducing the number of distinct individuals sequenced to 19 (10 females and 9 males) but providing an opportunity to estimate genotyping error (see *Variant Calling and Filtering: Variant Quality*). In addition, we sequenced a zebra finch domesticated family bred in captivity at East Carolina University, consisting of a mother, father, and three sons (Table S1). For long-tailed finch *Poephila acuticauda* (family: *Estrildidae*), we sequenced individuals from three populations in northern Australia, sampling ten individuals from each of the two long-tailed finch subspecies, *P. a. hecki* and *P. a. acuticauda* (8 females and 12 males; further details in Table S1 and (43). For use as an outgroup, we sequenced a single male individual of the double-barred finch *T. bichenovii*; this bird was part of a private individual's avifauna collection.

1.2 Reference Genome

We used the zebra finch reference genome originally published in August 2008 and commonly referred to as assembly taeGut1 or WUGSC 3.2.4 (`ftp://hgdownload.cse.ucsc.edu/goldenPath/taeGut1/chromosomes/`). This genome includes 1.0 Gb of the expected 1.3 Gb of sequence assembled in 31 autosomal chromosomes, three autosomal linkage groups, the Z chromosome, and the mitochondrial genome, as well as 200 Mb of additional unplaced sequence (19). In our analyses, we removed chromosome 1B because sequencing of other avian genomes (24) suggests that this chromosome was assembled in error.

This genome assembly includes the pseudoautosomal region (PAR), though it was not assembled to chromosome Z (see *Identifying the Pseudoautosomal Region*), but does not include the female sex chromosome W. We used Ensembl gene annotations, further curated for accuracy by the Avian Phylogenomics Consortium (available at `ftp://climb.genomics.cn/pub/10.5524/100001_101000/101000/zebrafinch/`). For all analyses using transcription start sites (TSSs) and end sites (TESs), we used the TSSs and TESs inferred in these annotations, which all fell at the starting codon and final codon, respectively. This suggests that 5′ and 3′ untranslated regions (UTRs) remain unannotated in the zebra finch; however, in birds, the length of 5′ UTR averages 100-200 bp and the length of 3′ UTR averages 650 bp (44). Thus, the unannotated UTRs should not affect our conclusions, given the kilobase scale at which we observe these patterns. For all analyses using CpG islands (CGIs), we used CGIs as annotated on the zebra finch genome CGI track on the University of California Santa Cruz Genome Browser. We used the same CGIs for zebra finch and long-tailed finch. We further note that experimental data suggest that CGIs can act as sites of transcriptional initiation, even when they are far from an annotated TSS (30), and thus, might be TSSs themselves. Thus, in all analyses, we use annotated TSSs to refer to only those TSSs that occur at the start of genes.

1.3 Species Tree Inference and TreeMix

To infer the species tree for the four species in this study, we first randomly selected 1000 loci of 1 kb length from across the 17 largest autosomes, selecting only those loci for which >80% of the bases were unmasked across all species, including our two outgroups, the medium ground finch (*Geospiza fortis*) and the collared flycatcher (*Ficedula albicollis*). To ensure reasonable run times, we randomly sampled four haplotypes from zebra finch and long-tailed finch and used both available haplotypes for the remaining three species. For each alignment, we identified the best-fit nucleotide substitution model using AICc scores estimated with MrAIC (45) and inferred the most likely gene tree using PhyML (46).

We then inferred species trees using three methods. First, taking all 1000 gene trees, we used STEAC to infer the species tree (47). Second, again using all 1000 gene trees, we used STAR to infer the species tree (47). Finally, for three randomly chosen subsets of the loci, each containing 50 loci, we inferred the species tree using a strict clock model linked across all loci in *BEAST (48). The clock rate for the analysis was set to $2.45 \cdot 10^{-3}$ substitutions / site / year based on our estimate of $7 \cdot 10^{-10}$ mutations / site / generation (see *Estimation of Mutation Rate*) and a zebra finch generation time of 3 - 4 generations per year (49). Each locus was assigned to the best-fit nucleotide substitution model inferred using MrAIC. Each run of *BEAST was run for $1 \cdot 10^{8}$ steps, sampling every $1 \cdot 10^{5}$ steps. Species tree topologies were congruent across all three methods, and divergence times for the three *BEAST runs were congruent as well. These divergence times inferred here are significantly earlier than previously published times for this group (50; 51; 52), which were based on fewer loci. However, these estimates accord well with estimates published more recently based on complete genomes (53). Despite this uncertainty, our estimates represent a lower bound on how deeply the conservation of hotspots extends.

In Figure 1, we report the *BEAST tree along with the 1000 gene trees inferred with PhyML, plotted using DensiTree (48), and present the same tree in Figure 4.

Further, to determine if these species exchanged migrants as they diverged, we used the program TreeMix (54). TreeMix was run for a random 20% of total autosomal SNPs across zebra finch, the two subspecies of long-tailed finch, and double-barred finch. We used the four-population test to determine if any inferred migration weights were significant (55).

Scripts used in this analysis are:

- `https://github.com/singhal/postdoc/blob/master/gene_trees.py`
- `https://github.com/singhal/postdoc/blob/master/convert_fasta_to_phyml.py`
- `https://github.com/singhal/postdoc/blob/master/make_treemix_input.py`

1.4 Confirming the Absence of PRDM9

To assess whether an ortholog of PRDM9 is indeed absent in birds, we used the human PRDM9 protein sequence to search for a PRDM9 ortholog in the genomes of 48 birds, one lizard, two turtles, and three crocodilian species (see Fig. S1 for full listing of species names and phylogeny; (53; 56). We first mapped the human PRDM9 protein sequence to the avian or reptile genomes with TBLASTN v2.2.23 (57), and excised the target-gene regions for gene structure and protein sequence determination using GeneWise v2-2-0 (58). This approach indicated that all 48 birds, the three crocodiles, and the lizard only have matches to the ZF domain of PRDM9 and have no

evidence for a full PRDM9 ortholog, i.e., all three domains of PRDM9 (ZF, SET, KRAB; (59)). In contrast, the two turtles sequenced, which are sister to all birds and crocodiles, appear to have a match that contains all three domains of PRDM9. Using BLASTP v2.2.23, we then mapped each of the protein sequences for the best matches back to the human gene set, and found that, except for the two turtles and two of the crocodilians, the best match for the predicted PRDM9 protein sequences to the human gene set was not back to human PRDM9. Together, these analyses suggest that PRDM9 is absent from all the bird species, but may be present in turtles and possibly some crocodiles. We additionally checked for PRDM9 in birds by using blat (60) to search PRDM9 amino acid sequences from five species (human, chimpanzee, cow, rat and mouse) against five bird species (budgerigar, chicken, medium ground finch, turkey and zebra finch) and found only short alignments. We also looked for an orthologous match in testes transcriptome data for zebra finch (see *Analysis of Expression Data*) and found that no candidate for PRDM9 was expressed at a detectable level.

1.5 Identifying the Pseudoautosomal Region

The pseudoautosomal region (PAR) in zebra finch was not annotated in the publication of the zebra finch genome. We therefore used sequence coverage data to predict the likely location of the PAR. To this end, we considered the ratio of coverage in females (ZW) on the Z chromosome, normalized by autosomal coverage, to the coverage in males (ZZ) on the Z chromosome, also normalized by autosomal coverage. Any region that is sex-linked should have a ratio of on average 0.5 because females are haploid for these regions. In contrast, for the PAR, this ratio should be on average 1. We estimated this ratio in 10 kb bins, removing repeats, across both chromosome Z and the unplaced chromosome Z_random, using coverage data calculated with bedtools genomeCoverageBed (61) and the realigned BAM files for both zebra finch and long-tailed finch (62). These results identified the PAR as the first 450 kb in chromosome Z_random, which is comparable in length to PARs identified in other, closely-related bird species (63). We confirmed our identification of the PAR by aligning the region identified in zebra finch to the PAR identified in collared flycatcher (64) and the medium ground finch (63) using LASTZ (65) under default settings (Fig. S24).

1.6 Sample Preparation and Sequencing

For each zebra finch unrelated sample, we extracted DNA from blood using Qiagen DNeasy Blood and Tissue Kit (cat no. 69504), and for long-tailed finch and double-barred finch, blood samples were extracted using a Qiagen PureGene kit (cat no. 158667). For all samples, we sheared DNA using Covaris set to 350 bp for 60s burst, yielding insert sizes of 320 bp on average, and constructed barcoded Illumina libraries using New England Biolabs Next DNA kit (cat no. E6040B). Following quantification with Invitrogen Qubit and quality assessment for sizing using Lab901 Tapestation, we pooled equimolar amounts of the samples across each species. Species pooled libraries were then diluted to a 10nM pool, and long-tailed finch and zebra finch libraries were sequenced across 15 lanes each using 100 bp paired-end reads with an Illumina HiSeq 2000. This work was performed by the Oxford Genomics Centre at the Wellcome Trust Centre for Human Genetics, Oxford, UK in 2012.

For the family of five zebra finches, DNA was extracted from liver tissue using Qiagen DNeasy Blood and Tissue Kit. Libraries were constructed and samples were barcoded and pooled together, then sequenced across six lanes on 2-lane flow cells of an Illumina HiSeq using 100 bp paired end

reads. This work was performed by the University of Chicago Genomics Core, Chicago IL USA in 2013.

1.7 Read Processing and Alignment

Following de-indexing of raw reads, reads from each individual were mapped to the zebra finch reference genome. All three species were mapped to zebra finch because there is no reference genome for long-tailed finch or double-barred finch and sequence divergence between the three species is low enough that almost all reads can still be reliably mapped (66). To accommodate species divergence and high genetic diversity within species, we used a sensitive aligner that would perform well even with read mismatches to the reference. For all wild-caught individuals, we ran Stampy v1.0.23 in hybrid mode, in which initial alignments are done with BWA and poorly-mapped reads then refined by Stampy (66; 67). For the domesticated zebra finch family, we used BWAmem v0.7.7 to map reads, because it has similar mapping success to the Stampy pipeline but has shorter runtimes (67). This approach worked well, with only small differences in alignment success across species; we recovered 97.6%, 96.6%, and 95.9% alignment rates for zebra finch, long-tailed finch, and double-barred finch, respectively.

Following initial alignment, we removed duplicates by individual using MarkDuplicates in Picard Tools v1.115, conducted local re-alignment around indels using RealignerTargetCreator and IndelRealigner in GATK v3.1-1 (62), and identified errors in mate-pairs using FixMateInformation in Picard Tools (`http://broadinstitute.github.io/picard/`). Representative commands for each step are available in *Representative Commands*.

1.8 Variant Calling and Filtering

1.8.1 Variant Calling

Variant calling was done separately for wild zebra finch, domesticated zebra finch, double-barred finch, and long-tailed finch for the 34 assembled chromosomes and linkage groups using the GATK UnifiedGenotyper pipeline (62). Our first step was to generate initial calls for both SNPs and indels and then to use this initial call set as known variants to input for base quality score recalibration (BQSR).

Next, the GATK best practices for variant calling suggests users provide a trusted set of variants to use as a training set for the variant score quality recalibration (VQSR) step. Because zebra finch has limited genomic resources, we were unable to use a "gold standard" existing variant set for calibration. Instead, we generated one ourselves. To this end, we first generated a call set for both SNPs and indels using the recalibrated alignments. Then, we used Cortex to generate a second call set; Cortex discovers variants by doing de novo assemblies and comparing the resulting assemblies to a reference genome (68). We found that assembling all individuals simultaneously with Cortex led to very few variants called, so we instead called each individual separately with Cortex using kmer=31 and then merged calls to generate a final Cortex call set. We took the intersection of the Cortex and the initial GATK call set as our trusted variants to use for training. Indeed, we found that using the intersected call set to recalibrate our variants led to fewer variants that were Mendelian errors in the domesticated zebra finch family than using either call set on its own (see *Variant Calling and Filtering: Mendelian Errors*). We used this training set to sequentially call and recalibrate SNPs and then indels via VQSR, retaining only those variants that passed VQSR.

Variants were further filtered (see *Sex Chromosomes*, *Coverage and Repeat Masking Filtering*, *Mendelian Errors*, and *Phasing*) to generate final call sets. Details on number of variants discovered can be found in Table S2.

Representative commands for each step are available in *Representative Commands*.

1.8.2 Sex Chromosomes

The UnifiedGenotyper module in GATK has no built-in support for differing ploidy levels for sex chromosomes in males and females. Thus, we modified variant calls to account for ploidy using in-house scripts. The pseudo-autosomal region (PAR) does not appear to be on the Z-chromosome as assembled in taeGut1, so we expect females to be haploid at all variants on the assembled Z. For any variants where a female was called as heterozygous for a given variant, we recoded the genotype as missing. 165,422 genotypes (1.8%) in zebra finch and 233,357 genotypes (1.5%) in long-tailed finch were recoded as missing. If three or more females were heterozygous for a given SNP, we suspected that the SNP was called in error and filtered the SNP from our data set for all individuals. In total, 16,513 Z-linked sites (2.0%) in zebra finch and 30,263 (1.6%) sites in long-tailed finch were filtered using this cutoff.

The script used in this work: `https://github.com/singhal/postdoc/blob/master/chrZ_postprocess_vcfs.py`.

1.8.3 Coverage and Repeat Masking Filtering

Regions with exceptionally low or high coverage often indicate duplicated regions that were collapsed in the genome assembly or other forms of mismapping. Given this concern, we first determined the coverage at each base in the genome for each individual using samtools depth v0.1.19 (69). From these data, we calculated the average sequencing coverage across the reference assembly for all individuals of a given species. We then filtered variants that were at bases with lower than 0.5× or higher than 2× coverage than the average (70).

We further filtered the variant call sets by removing any sites that occurred in repeat-masked regions, using repeat annotation for zebra finch from RepeatMasker Repeat Library 20140131, downloaded on 24 October 2014 from RepeatMasker (`http://www.repeatmasker.org/species/taeGut.html`).

The scripts used in this work are:

- `https://github.com/singhal/postdoc/blob/master/make_masked_genome.py`
- `https://github.com/singhal/postdoc/blob/master/make_vcf_filtered_for_coverage_repeatmasked_by_chr.py`

1.8.4 Mendelian Errors

Using the five-individual zebra finch family, we used PLINK v1.07 to identify variants that showed evidence for Mendelian errors (71). These sites were then filtered from the zebra finch variant set (see *Variant Quality*). A representative PLINK command is available in *Representative Commands*.

1.8.5 Variant Quality

We calculated several statistics to better understand the quality of our variant data. First, one individual in the wild zebra finch population was sequenced twice and genotyped independently, allowing us to compare genotype concordance across two replicates for the same biological sample. We found that 99.57% of autosomal genotypes were concordant, 0.13% were missing in one of the replicates, and 0.30% were discordant. For genotypes on chromosome Z, we found that 98.81% genotypes were concordant, 0.67% were missing in one of the replicates, and 0.52% were discordant. Second, we calculated the Mendelian error rate within the domesticated zebra finch family, finding an error rate of 0.36% per SNP and 0.79% per indel for autosomes. Third, because females are haploid for the Z chromosome, if females are heterozygous for any site on the Z chromosome, that variant is in error. Considering the number of SNPs that are called as heterozygous in females over the total number of SNPs on the Z suggests a false discovery rate (FDR) in females of 3.8% in zebra finch and 2.5% in long-tailed finch. In practice, error rates are likely somewhat higher, as even hemizygote sites may be miscalled. While these values are comparable to those reported in other studies (9), given the apparent lower data quality of the Z, we analyze this chromosome separately.

Moreover, we note that a number of other features of the Z chromosome make inferences based on linkage disequilibrium harder to interpret: the presence of long inversions (32; 33); the high F_{ST} between subspecies of long-tailed finch (F_{ST}=0.17); and the greater variation of diversity levels with recombination (Fig. S27C-D), which is potentially indicative of stronger effects of linked selection. In particular, these Z-linked inversions could reduce recombination rates on the Z by up to a third but should not depress estimated rates further unless they are much older than typical polymorphisms and so have built up linkage disequilibrium (72). If the inversion were unusually old, however, diversity levels between karyotypes would be unusually elevated, when diversity is in fact lower than on comparably sized autosomes. Given caveats regarding data quality and the unusual patterns of variation seen on chromosome Z, we consider analyses for chromosome Z with caution.

We note further that polymorphisms segregating at low frequency (i.e., singletons) likely have higher error rates (9; 70). Importantly, all singletons were removed before we inferred the recombination map and identified hotspots.

1.8.6 Ancestral Allele Inference

We inferred the ancestral allele for all variable SNPs in either zebra finch or long-tailed finch so that we could polarize all mutations. To do so, we used phylogenetic data from two other sequenced finch species, the medium ground finch and the large ground finch, *G. magnistrosis* (23; 73), which share a root with our focal species about 15 million years ago (Fig. 4), and applied a simple parsimony approach.

For each species, we downloaded the raw short read sequence data from the National Center for Biotechnology Information Short Read Archive (medium ground finch: PRJNA156703; large ground finch: PRJNA178982), aligned the reads to the zebra finch genome using Bowtie2 v2.2.23 (74) and identified variable sites using samtools (69). Alignment rates for the medium and large ground finch to the zebra finch genome were 82.6% for 431 million read pairs and 86.4% for 11.3 million read pairs, respectively, suggesting that these species' genomes are sufficiently similar to allow mapping across these phylogenetic distances. Example command lines for generating these

genomes are included in *Representative Commands*.

Then, for each given variable site, we looked at the identity of the base in the other species; if all the other species had the same identity and if that identity was one of the segregating alleles at the variable site, we called that the ancestral allele. If multiple species were polymorphic for a given site, we only considered those alleles that were fixed in the other species and called the most common allele as the ancestral allele, requiring more than half of the species to have the most common allele. For sites where this approach did not resolve the ancestral allele, we set the ancestral allele to 'N', or unknown.

Scripts used in this work include:

- `https://github.com/singhal/postdoc/blob/master/simple_ancestral_chromosomes.py`
- `https://github.com/singhal/postdoc/blob/master/geospiza_genomes.py`
- `https://github.com/singhal/postdoc/blob/master/simple_darwin_ancestral.py`

1.8.7 Estimation of Nucleotide Diversity and the Mutation Rate

To estimate nucleotide diversity, we used all SNPs, including those that are multiallelic, to calculate Watterson's θ (75) and π (76). Our resulting estimates of π=0.082 in zebrafinch are similar to that obtained surveying a much broader geographic sampling with a small number of loci (π=0.01, (77)).

To obtain a rough estimate of the mutation rate, we used the relationship between ρ and θ, $\frac{4N_e\mu}{4N_ec} = \frac{\theta}{\rho}$. Because our simulation results showed that ρ/bp estimates tend to be less accurate when the background ρ/bp is higher than 0.8, we estimated π and θ for the 17 autosomal chromosomes below this cutoff. We estimated Watterson's θ by counting the number of segregating SNPs (75), estimated ρ by calculating the median ρ value inferred in the maps generated under block penalty 100, and estimated c by finding the median value given in Backstrom et al. (21). This calculation yields an inferred mutation rate of 7 $\cdot 10^{-10}$ mutations / site / generation, or 2.1 - 2.8 $\cdot 10^{-9}$ mutations / site / year, given 3 to 4 generations per year in zebra finch (49), which accords well with fossil-based estimates of the substitution rate in *Passeriformes* (2.8 - 3.6 $\cdot 10^{-9}$ / site / year) (41).

1.9 Phasing Haplotypes

1.9.1 Phasing

There are several approaches for computationally phasing individuals, which include using patterns of linkage disequilibrium, pedigree data, and phase-informative reads to infer haplotypes from genotypes. Given the difficulties in accurately phasing data, we used all three sources of information to phase our data. First, we phased the five-individual zebra finch family using HAPI v1.03 in "mr mode" (78), which uses pedigree data to infer haplotypes with the fewest recombination events (see *Representative Commands*). We then used the results to identify regions in which phasing was low quality. Specifically, by comparing inferred offspring haplotypes to parental haplotypes, we can call putative recombination breakpoints. In doing so, we found that many breakpoints were very close together (i.e., 92.2% spanned less than ten heterozygous sites). While a subset of these events could represent non-crossover resolutions, most are more likely due to

sequencing or mapping artifacts. Thus, we masked the region between such breakpoints from the zebra finch variant sets. After masking these regions, we used HAPI to again phase the haplotypes of the zebra finch family and used the four parental haplotypes as a reference set of haplotypes in the next step.

Specifically, we used the four parental haplotypes, phase-informative reads, and patterns of linkage disequilibrium to guide phasing of the zebra finch individuals, as implemented in Shapelt's assemble mode (v2r790) (79). Briefly, Shapelt uses alignment data for each individual to identify sequencing reads that span two heterozygous sites. Phase information from such reads is then used in combination with a linkage disequilibrium model to phase all sites for all individuals. Shapelt in assemble mode can only phase biallelic SNPs, and thus, we attempted to phase multiallelic sites by recoding them as two overlapping biallelic SNPs. However, this approach lead to segmentation faults when running Shapelt, so we removed all multiallelic sites from the analysis (1,643,215 (3.7%) of sites in zebra finch, 465,636 (1.8%) of sites in long-tailed finch). Representative commands for this approach are included in *Representative Commands*.

To phase the Z chromosomes, we modified this approach by including the known haplotypes from females as a reference set of haplotypes when we phased the males.

For long-tailed finch, because we did not sequence any pedigreed individuals, we phased the individuals solely using phase-informative reads and linkage disequilibrium information in Shapelt's assemble mode.

Scripts used in this work include:

- `https://github.com/singhal/postdoc/blob/master/recombination_breaks_hapi.py`
- `https://github.com/singhal/postdoc/blob/master/find_recombination_breaks_across_lengths_hapi.py`
- `https://github.com/singhal/postdoc/blob/master/switch_error_rate_compare_hapi_PIR_family.py`
- `https://github.com/singhal/postdoc/blob/master/filter_vcf_for_switch_errors.py`

1.9.2 Phasing Errors

Errors in phase, typically measured as a switch error rate (80), can be large, particularly when phasing a relatively small number of unrelated individuals. Because phasing errors can artificially elevate estimates of recombination rates, it is important to estimate phasing error rates. To that end, we estimated the average level of switch error rate across the genome by comparing haplotypes phased by two approaches. First, we phased the zebra finch family parents using HAPI v1.03 as described above. Second, we re-phased the 19 unrelated zebra finch individuals along with the two zebra finch parents using Shapelt as described above, but omitting the reference haplotypes. Assuming that family phasing is nearly perfect, we can then infer the switch error rate by comparing the haplotypes of the parents as phased by HAPI and as phased by Shapelt. Here, we calculated switch error rate excluding singletons, as singleton sites were not used to infer recombination maps. As averaged across the four parental haplotypes, we inferred a median error rate of 1.2%-8.9% across long chromosomes in zebra finch (Fig. S29); for comparison, using similar approaches to phase as used here with much larger populations in humans results in error rates of $\sim$0.5% to 1.0% (79). Importantly, we note that our inferred hotspots do not have higher switch error rates than do other regions (Fig. S9C).

1.10 Fine-Scale Recombination Maps

1.10.1 Generating Recombination Maps

To generate recombination maps, we used a linkage-disequilibrium approach that infers recombination rates from haplotype estimation. While this approach only generates a historical, sex-averaged map of recombination (81), other methods such as sperm typing, double strand break sequencing, or analyses of pedigrees are impractical for most non-model species (82).

To generate fine scale estimates of recombination rate variation across the chromosomes, we used the program LDhelmet v1.6 (15). In simulation studies, LDhelmet shows improvement in estimation accuracy compared to LDhat (83). To infer recombination rates, LDhelmet requires: (1) a set of phased haplotypes, from which we removed singletons because they cannot be reliably phased and are almost uninformative about recombination; (2) an estimate of θ, the population mutation rate, for the samples being phased; (3) the ancestral allele at each variable site (see *Ancestral Allele Inference*); and (4) a mutation matrix (see *Mutation Matrix Estimation*). We estimated Watterson's θ based on the number of non-singleton biallelic SNPs segregating in each species (θ_w for zebra finch = 0.68%; for long-tailed finch = 0.47%). Representative commands for running LDhelmet are included in *Representative Commands*.

A key LDhelmet parameter is the block penalty, which defines the penalty incurred to the likelihood every time the recombination rate changes across the genome. Higher block penalties result in smoother recombination maps. Without prior knowledge of the fine-scale recombination landscape of the species, there is no obvious choice of block penalty for a data set, so we conducted a series of simulations to define the block penalty (see *Defining Block Penalty*). These simulations suggested that, for plausible parameters for these species, a block penalty of 5 gives the most power to detect hotspots whereas a block penalty of 100 provides the most accurate recombination maps across megabases of sequence (see *Defining Block Penalty*; Figs. S30, S31). Thus, we ran each chromosome at both block penalties for $1 \cdot 10^6$ steps in the MCMC chain with a burn-in of $1 \cdot 10^5$ steps.

The script used in this work was `https://github.com/singhal/postdoc/blob/master/make_seq_for_ldhelmet.py`.

1.10.2 Ancestral Allele for LDhelmet

LDhelmet requires ancestral alleles to be defined, so we used the ancestral states that we had inferred using parsimony (see *Ancestral Allele Inference*). Because LDhelmet allows users to provide a prior probability for each base as the ancestral allele, to allow for uncertainty in ancestral allele reconstruction, we set the prior for the putative ancestral allele as 0.91 and the other three bases' priors as 0.03. For sites where we could not resolve the ancestral allele, we indicated this uncertainty in LDhelmet by providing prior probabilities for each nucleotide that were equal to their stationary frequency, as estimated from empirical frequencies in the genome and the mutation matrix (see *Mutation Matrix Estimation*).

1.10.3 Mutation Matrix Estimation

To estimate the mutation matrix for zebra finch and long-tailed finch, we first considered only biallelic SNPs. For those SNPs for which we could infer the ancestral allele (see *Ancestral Allele*

Inference), we determined the mutation type, resulting in a 4x4 matrix of counts of inferred mutation types (Table S6). We then followed the approach outlined by (15) to estimate the normalized mutation matrix for each species.

The script used for this inference is `https://github.com/singhal/postdoc/blob/master/get_mutation_matrix.py`.

1.10.4 Defining the Block Penalty

To define the appropriate block penalty to use for both estimating the recombination map and identifying putative hotspots, we conducted a series of simulations. These simulations were designed to mimic the characteristics of the zebra finch dataset, though we note that zebra finch and long-tailed finch are broadly similar in terms of the range of recombination rates seen, estimates of θ, and number of haplotypes sampled. For these simulations, we used MAcS v0.5d (84) to simulate chromosomal segments of 1 Mb each for 38 haplotypes, using the θ (with singletons) and mutation matrix inferred for zebra finch. We repeated each simulation for a series of background recombination rates (i.e., rates outside of hotspots) that spanned the full range of ρ/bp values seen in preliminary recombination maps: 0.0001, 0.001, 0.01, 0.1, 1, and 2.5. For each ρ value, we placed eight hotspots throughout the 1 Mb of sequence, two each for hotspots of relative heat 10×, 20×, 40×, and 60×. Each parameter set was repeated 12 times, giving a total of 24 hotspots simulated for each ρ and each relative heat. Then, after removing singletons from the simulated haplotypes, we used LDhelmet to estimate the recombination rate for these sequences, using for each simulation a range of block penalties of 5, 10, 50, 100, and 500.

For these simulation results, we asked two questions. First, which block penalty provides the most accurate general picture of recombination rates? To address this question, we calculated how much inferred rates from LDhelmet deviated from known recombination rates across the entire length of the 1 Mb of simulated sequence. These results suggested that block penalty 100 provided the most accurate estimation of background recombination rate across a range of recombination rates (Fig. S30). Second, we asked which block penalty provides the best power to identify hotspots. To answer this question, we identified putative hotspots in the inferred recombination maps, identifying regions 2 kb or greater that had 5× or greater recombination rate than their 80 kb of surrounding sequence, as was done with the empirical data. From this, we calculated the power to identify hotspots at each given parameter set, finding that block penalty 5 provided the most power (Fig. S31).

We then ran a second set of simulation results using the same basic format outlined above, but allowing θ to vary across the genome-wide range of values seen in zebra finch. These simulations aimed to see if θ and ρ interact to influence power to detect hotspot power. We ran the simulations for three θ/bp values (0.0075, 0.014, 0.02) and five ρ values that varied at 0.01×, 0.1×, 1×, 10×, 100× fold the θ value. For each θ and ρ value, we ran the simulation ten times, resulting in 25 hotspots at each given parameter set. This work showed that power was a function of both θ and ρ, and not simply their ratio (Fig. S32).

The scripts used in this work:

- `https://github.com/singhal/postdoc/blob/master/simulations_hotspot_power.py`
- `https://github.com/singhal/postdoc/blob/master/simulations_hotspot_power_theta.py`

- `https://github.com/singhal/postdoc/blob/master/simulations_find_hotspots1.py`
- `https://github.com/singhal/postdoc/blob/master/simulations_find_hotspots2.py`

1.10.5 Comparison to an Existing Genetic Map

A linkage map for zebra finch was previously inferred by genotyping ∼1920 SNPs in a multi-generation pedigree of approximately 1000 birds in Backstrom et al. 2009, providing resolution at the scale of tens of megabases (21). To compare their genetic map to the one inferred in this study, we used the framework map that Backstrom et al. published, which only includes SNPs whose LOD score was greater than 3 and discards outliers (defined as any loci whose genetic location relative to others in the map did not correspond to its physical location). This framework map consists of 443 SNPs; after removing loci for which we did not have physical positions, we retained 437 SNPs. We converted ρ in both the zebra finch and long-tailed finch maps to cM/Mb by scaling each chromosome by the genetic map length for the same chromosome in the Backstrom et al. map. Initial explorations of the data showed that while the shapes of the maps inferred by Backstrom et al. and those we inferred were concordant, if we assumed the per chromosome genetic lengths between the two maps were the same, the maps appeared highly discordant. Instead, we allowed the total genetic length per chromosome to vary between the Backstrom et al. map and our map, and scaled each chromosome by a scalar that we obtained by finding the least-squares fit between the map inferred from our data and that inferred by Backstrom et al. Doing so, we found that the total genetic length of each chromosome in our map was, on average, 14.9% longer than in the Backstrom et al. map. This difference is potentially consistent with the limited number of markers in the pedigree study.

1.11 Hotspots

1.11.1 Power to Detect Hotspots

To assess our power to detect hotspots, we conducted a series of simulations, again mimicking the characteristics of the zebra finch dataset in terms of θ and ρ estimates, the number of haplotypes, and the mutation matrix. For these simulations, we used MAcS to simulate chromosomal segments of 1 Mb each for 38 haplotypes (84). We repeated each simulation for a series of background recombination rate (i.e., rates outside of hotspots) that spanned the full range of ρ/bp values seen in preliminary recombination maps: 0.0001, 0.001, 0.01, 0.1, 0.2, 0.4, 0.6, 0.8, 1, and 2.5. For each ρ value, we placed 8 hotspots throughout the 1 Mb of sequence, two each for hotspots of relative heat 5×, 10×, 20×, and 40×. Each parameter set was repeated 50 times, giving a total of 100 hotspots simulated for each ρ and each relative heat. We used LDhelmet to estimate the recombination rate for these sequences at the same block penalty used for recombination maps for hotspots, block penalty 5. We then calculated how many hotspots could be inferred out of the simulated hotspots, identifying regions 2 kb or greater that had 5× or greater recombination rate than the background rate, as calculated across 40 kb flanks. We note that, for our actual data, we validated our hotspots in a second step, using sequenceLDhot (see *Hotspots: Identifying Hotspots*) (85). This validation step likely reduces our power to identify hotspots, so the power we report here is likely an over-estimate.

1.11.2 Identifying Hotspots

Our simulation results showed that we have little power to identify hotspots when the background ρ is below 0.0001/bp or exceeds 0.8/bp. Thus, we limit our detection of hotspots to those 18 chromosomes for which the median background ρ is within 0.0001/bp and 0.8/bp for zebra finch and long-tailed finch: chromosomes 1, 1A, 2, 3, 4, 4A, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and Z. For these chromosomes, we first identified putative hotspots by calculating hotspot ρ for 2 kb central windows, slid by 1 kb every iteration, and background ρ for 40 kb flanks on either side using recombination maps generated by LDhelmet under block penalty 5. Putative hotspots were any windows whose ρ was at least 5-fold greater than that of flanking regions.

After defining putative hotspots in both species, we filtered the list of putative hotspots to remove any hotspots that occurred within 5 kb of another putative hotspot and then validated these hotspots using sequenceLDhot (85). For windows across a given sequence, sequenceLDhot compares the relative likelihood of the data under a model in which window ρ is higher than background ρ versus a model in which the window ρ and background ρ are equal. We ran this test on a 50 kb region centered on the putative hotspot location, again removing singleton sites from the haplotypes, and providing the background ρ as inferred by LDhelmet. For a hotspot to be validated, we required it to be supported with a likelihood ratio cutoff of $\geq$10 and required the estimated heat to be $\geq$5.

We classified hotspots between zebra finch and long-tailed finch as shared when their midpoints occurred within 3 kb of each other. We also explored how different criteria for sharing impacts the percentage of hotspots shared, finding that our estimate of hotspot sharing is robust to the use of different criteria (Table S3).

Scripts used in this work included:

- `https://github.com/singhal/postdoc/blob/master/find_hotspots.py`
- `https://github.com/singhal/postdoc/blob/master/find_hotspots_parse.py`
- `https://github.com/singhal/postdoc/blob/master/find_hotspots_seqldhot_validate.py`
- `https://github.com/singhal/postdoc/blob/master/find_hotspots_seqldhot_parse.py`

1.11.3 Hotspot Validation

We further profiled the validated hotspots to confirm that our findings did not arise from methodological artifacts or quality control issues. First, we compared data quality in hotspot regions to non-hotspot regions, measuring data quality by rates of switch error, numbers of Mendelian errors, percentage of data masked for coverage and repeats, and quality of SNPs. These comparisons showed that data quality at hotspots was, if anything, slightly better than at non-hotspot sequences (Fig. S9). Second, we took advantage of Shapelt's reporting of the haplotype graph, which encapsulates uncertainty in phasing. From the haplotype graph, we drew three samples of alternate, likely phasings for a subset of validated haplotypes and determined if the hotspots were validated under these alternate phasings as well (Table S5). Third, since sequenceLDhot requires background ρ to be defined a priori, we ran a subset of validated hotspots under both decreased (0.5$\times$ of estimated rate) and elevated (1.5$\times$) background ρ to determine how misspecification of

this parameter might influence the likelihood of validating a hotspot (Table S5). All these analyses lent support to the reliability of our hotspot inferences. Further validation is provided by the evidence for localized biased gene conversion (see *Estimation of GC**).

One caveat concerns hotspots inferred to exist in only one species. Simulations across a range of background recombination rates and hotspot heats indicate that the observed levels of hotspot sharing are somewhat lower then expected under a model in which all hotspots are identical in heat and location (Fig. S12). This finding suggests that beyond incomplete power, there could be true differences in a subset of hotspots between the two species. To investigate this possibility, we considered the small number of cases (202 in ZF and 332 in LTF) in which we had statistical support for a hotspot in one species and no support in the other (i.e., for which sequenceLDhot inferred hotspot heat to be equal to 1 and the likelihood ratio test was less than or equal to 1). Although we have sufficient power to detect the effects of gBGC when we subsampled similar numbers of hotspots from the total set, these so-called "unique hotspots" show no evidence for gBGC in either species (Fig. S33A-B). However, for both species, these unique hotspots did show evidence for biased mutational spectra (as measured by derived allele frequency) in both the species in which they were presumed to be unique and in the other species (Fig. S33C-D). Thus, these "unique hotspots" might actually be shared.

Scripts used in this work include:

- `https://github.com/singhal/postdoc/blob/master/find_hotspots_phasing_uncertainty.py`
- `https://github.com/singhal/postdoc/blob/master/find_hotspots_phasing_uncertainty_parse.py`
- `https://github.com/singhal/postdoc/blob/master/find_hotspots_seqldhot_check_sensitivity.py`

1.11.4 False Positive Rate

High switch error rates can artificially increase the inferred recombination rate, so we conducted a series of simulations to determine the incidence of spurious hotspot calls under varying levels of switch error rate. For these simulations, we used MAcS to simulate chromosomal segments of 1 Mb each for 38 haplotypes, using the same θ and mutation matrix inferred for zebra finch. We repeated each simulation for a series of background recombination rate that spanned the full range of ρ/bp values seen in long chromosomes: 0.001, 0.01, 0.1, and 0.5. For each ρ value, we simulated a range of switch error rates that spanned the full range seen: 0.0, 0.01, 0.02, 0.04, 0.08, and 0.16. Each parameter set was repeated 12 times. We then used LDhelmet to infer the recombination maps with a block penalty 5. To calculate the false positive rate, we determined how many, if any, hotspots were inferred. We found that the number of false positives inferred varied greatly depending on background recombination rates, and the magnitude of switch error rate only had an effect when background recombination rates were low (Fig. S34).

The script used for this analysis is available here: `https://github.com/singhal/postdoc/blob/master/simulations_FDR.py`.

1.11.5 Likelihood of Shared Hotspots

We define shared hotspots as those whose midpoints occur within 3 kb of each other. Initial results suggested that zebra finch and long-tailed finch shared a large percentage of hotspots ($\geq$70%). Given that we do not have perfect power, this observation is likely consistent with a greater percentage of shared hotspots. We therefore conducted a series of simulations to assess what percentage of hotspots would be inferred as shared if hotspots shared between the two species had the same locations and heats. To do so, we used MAcS to simulate two population samples of 1 Mb of sequence each at a range of background ρ/bp values seen in long chromosomes. Because the average ρ in zebra finch is approximately double that of long-tailed finch, the background ρ/bp for population 1 ranged from 0.001, 0.002, 0.01, 0.1, 0.5, and 0.8, whereas the corresponding ρ/bp in population 2 ranged from 0.0005, 0.001, 0.005, 0.05, 0.25, and 0.4. On each sequence, we placed 12 hotspots each, with hotspot heat varying from 5×, 10×, 15×, 20×, 40×, 60×, 80×, and 100×. The first population sample was simulated for 38 haplotypes, using the same θ and mutation matrix inferred for zebra finch, and the second was simulated for 40 haplotypes, using the same θ and mutation matrix inferred for long-tailed finch. Each simulation was repeated 10 times, giving a total of 120 hotspots simulated per parameter set in each species. After simulations, we used LDhelmet to infer recombination maps, identified putative hotspots, and used sequenceLDhot to validate in both species all putative hotspot locations identified in either species 1 or species 2. Then, after binning hotspots by their estimated background ρ and hotspot heat in species 1, we calculated the percentage of shared hotspots per bin.

We further estimated the rate at which we expect to see spurious cases of sharing. There are two possible cases where we could identify spurious sharing. In the first, the hotspot is a false positive in both species 1 and species 2. We identified no cases of spurious sharing by this definition. In the second case, the hotspot is real in either species 1 or species 2 but is a false positive in the other species. To model this scenario, we compared the data simulated for population 1 (as above) to data simulated for population 2, in which we simulated sequences of 1 Mb each with no hotspots, again with a ρ/bp that was half that of population 1 and the θ and mutation matrix for long-tailed finch. For each parameter set, we simulated 120 hotspots across 10 Mb sequence in population 1 and 10 Mb of sequence in population 2. In this scenario, we also identified no cases of spurious sharing. Thus, we conclude that the rate at which we would infer spurious sharing is negligible.

Scripts used in this analysis include:

- `https://github.com/singhal/postdoc/blob/master/simulations_shared_hotspots.py`
- `https://github.com/singhal/postdoc/blob/master/simulations_shared_hotspots_find_hotspots2.py`

1.11.6 Null models for Hotspot Sharing

To estimate the number of hotspots that would be inferred as shared simply by chance, we created a null model, which we applied to both zebra finch/long-tailed finch and chimpanzee/human pairs. Our empirical data suggest that our inferred hotspot locations are not randomly distributed across the genome; we tend to find fewer hotspots at low and high recombination regions of the genome than we would expect (Fig. S35). Given that, we created a null model for hotspot locations that conditions hotspot placement on the background recombination rate of the sequence. To do so, for each chromosome, we characterized the background recombination rate across 100 kb

windows, binning each window into deciles. We then determined in which windows we found hotspots, and from that, determined the empirical distribution of hotspots across recombination rate deciles. In simulating hotspots, we used this same empirical distribution to randomly place hotspots down across the chromosome. We repeated this for each species and then used these randomly placed hotspots to calculate the percent of hotspots shared under the null. In each comparison, percent of hotspots shared is calculated using the number of hotspots in the species with fewer hotspots as the denominator. We generated 1000 random sets of hotspots to generate a null distribution.

We used existing data on hotspots in humans and chimpanzees to generate a null expectation for hotspot sharing between these two species. Chimpanzee hotspots were downloaded in pantro2 coordinates from `ftp://birch.well.ox.ac.uk/panMap/haplotypes/genetic_map/hotspots/` (9). There were 5,038 hotspots, and all but 57 could be lifted over to hg18 using the liftOver tool from the UCSC Genome Browser. Human hotspots were downloaded in hg17 coordinates `ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2006-10_rel21_phaseI+II/hotspots/` (86). All 34,142 of these hotspots could be converted to hg18 coordinates. We also used existing data on recombination rates in syntenic regions between humans and chimpanzees to characterize background recombination rates (`ftp://birch.well.ox.ac.uk/panMap/haplotypes/syntenic_genetic_map/`), smoothing these raw data across 100 kb windows (9).

1.11.7 Hotspot conservation across the avian phylogeny

In order to gain a deeper time perspective on hotspot conservation across the avian phylogeny, we estimated GC* at hotspots (see *Estimation of GC**) for two other species of birds, the medium ground finch and collared flycatcher (87). We defined the genome sequence of medium ground finch with respect to the zebra finch genome as described in *Ancestral Allele Inference* and applied an identical approach to defining the genome sequence of the collared flycatcher. For the collared flycatcher, we used the reads in NCBI BioProject PRJEB2984, aligning the reads to zebra finch and calling variants using samtools (see *Ancestral Allele Inference* for details). Alignment rates were 72.9% across 565 million read pairs.

We redefined the ancestral genome sequence to include the medium ground finch and the collared flycatcher following the same algorithm outlined in *Ancestral Allele Inference*. The resulting ancestral and species-specific genome sequences were used to estimate GC* for the medium ground finch and the collared flycatcher.

1.11.8 Matched hotspots and coldspots

To better understand the unique properties of hotspots, we identified a set of coldspots that we could compare to the hotspots, matching the coldspots to the hotspots for GC and CpG content. Coldspots were defined as 2 kb windows in which ρ was 0.9 - 1.1-fold the background ρ, for which background ρ/bp varied from 0.001 to 0.1, and which occurred more than 25 kb from a hotspot. These criteria resulted in 31,016 coldspots in zebra finch and 49,741 coldspots in long-tailed finch.

We calculated the ancestral GC and CpG content in a 10 kb window centered on each hotspot or coldspot, using the ancestral sequence that we had inferred and excluding sites that had been masked in our pipeline for that species or for which the ancestral base was unknown. We then removed hotspots and coldspots for which more than half of the 10 kb window was excluded or that were less than 5 kb from the end of a chromosome. Next, we matched each remaining hotspot

to the closest coldspot located on the same chromosome that had GC content within 1% and CpG content within 0.1% that observed and that did not overlap with other matched coldspots. We were able to match all but 292 zebra finch hotspots and 199 long-tailed finch hotspots, giving us a set of 3,657 matched hotspots and coldspots in zebra finch and 4,734 in long-tailed finch.

For double-barred finch, medium ground finch, and collared flycatcher, we used the shared hotspots between zebra finch and long-tailed finch as a set of putative hotspots. For these, we identified matching coldspots, using coldspots identified in long-tailed finch, again requiring ancestral GC content to be within 1% and CpG content within 0.1% of that observed (we used long-tailed finch coldspots, as we believe rate estimates in long-tailed finch to be more reliable than those in zebra finch). We were able to match all but 104 hotspots for double-barred finch, all but 111 hotspots for medium ground finch, and all but 120 hotspots for collared flycatcher, giving us 2,770, 2,763, and 2,754 matched hotspots, respectively.

Scripts used in this analysis are:

- `https://github.com/singhal/postdoc/blob/master/identify_coldspots.py`
- `https://github.com/singhal/postdoc/blob/master/match_hotspots_and_coldspots.py`

1.11.9 Estimation of GC*

We estimated GC* (88; 89), the expected equilibrium GC content, in 100 bp windows from the center of the matched hotspots and coldspots, as:

$$GC^* = \frac{\frac{AT \rightarrow GC}{ancAT}}{\frac{AT \rightarrow GC}{ancAT} + \frac{GC \rightarrow AT}{ancGC}} \quad (1)$$

where ancAT and ancGC are the number of As and Ts or Gs and Cs, respectively, in the ancestral sequence, and AT→GC and GC→AT represent the number of substitutions on a lineage from an A or T to a G or C, or from a G or C to an A or T, respectively. Substitutions where either ancestral or derived allele creates a CpG site were excluded because of the difficulty in reconstructing the ancestral sequence at rapidly-evolving CpG sites. We inferred the number of substitutions separately on the zebra finch, long-tailed finch, double-barred finch, medium ground finch, and collared flycatcher branches by comparing the sequence data that we obtained for each species with the ancestral sequence. We considered sites where all sampled individuals from that species carry non-ancestral alleles and all sampled individuals from the other species carry ancestral alleles (i.e., fixed differences between species samples) as lineage-specific substitutions. Note that, although we did not calculate GC* for the large ground finch because it is so closely-related to the medium ground finch, we did include its sequence data in inferring ancestral alleles and identifying lineage-specific mutations in the other species.

To look for a difference in the frequency of GC alleles in zebra finch and long-tailed finch consistent with biased gene conversion, we calculated the mean frequency of the derived allele for each type of mutation (A or T to G or C, G or C to A or T, A or T to A or T and G or C to G or C) in 100 bp windows from the center of the matched hotpots and coldspots. We considered the derived allele as polarized by the ancestral sequence and again excluded sites where either allele creates a CpG site due to difficulty in reconstructing the ancestral sequence at these rapidly-evolving sites.

Scripts used in this work include:

- `https://github.com/singhal/postdoc/blob/master/gc_hotspots_darwin.py`

- https://github.com/singhal/postdoc/blob/master/calculate_gcstar.py

1.11.10 Motif Discovery

To discover motifs that were enriched in hotspots relative to coldspots in both zebra finch and long-tailed finch, we used the program MEME (90), which allows for discriminatory analysis of motifs when used with the program psp-gen. Because MEME has exponential increases in run times with increased number of sequences to search, we limited our motif discovery to 1000 randomly selected hotspots occurring on autosomal chromosomes and their matched coldspots and ran MEME on 1 kb regions surrounding the center of the spot, for the ancestral sequence. When running MEME, we allowed motif size to vary from 5-mers to 20-mers and searched on both strands for motifs. We ran this analysis five times to ensure the results could be replicated. Representative commands used to run these analyses are available in *Representative Commands*.

1.12 Analysis of Gene Expression

We used RNAseq to estimate gene expression in the testes for six zebra finch males. Prior to this experiment, these birds were held in captivity at the University of Illinois. They were socially isolated overnight (for unrelated experiments) and euthanized the next morning. All procedures were conducted under protocols approved by the University of Illinois Institutional Animal Care and Use Committee.

Testes were snap frozen on dry ice. The testes tissue was homogenized in Tri-Reagent (Molecular Research Company) and total RNA was extracted following manufacturer's instructions. Total RNA was then DNase treated (Qiagen) to remove any genomic DNA contamination and the resulting RNA was purified using Qiagen RNeasy columns. Total RNA was assessed for quality using an Agilent Bioanalyzer. Library preparation and sequencing were done at the University of Illinois Roy J. Carver Biotechnology Center following Illumina TruSeq RNA Sample Prep Kit and manufacturer's protocols, and the libraries were sequenced on an Illumina HiSeq 2000 using a TruSeq SBS sequencing kit version 3 producing single end reads which were analyzed with Casava 1.8.2.

Reads were aligned to the zebra finch reference genome using bwa-mem (v. 0.7.10-r789) with default settings (67). Expression data for zebra finch genes, as catalogued in Ensembl 74 (91), was calculated using eXpress v. 1.5.1 under default settings (92). These analyses yielded estimates of gene expression, measured as fragments per kilobase of exon per million reads mapped (FPKM), for 9,281 genes.

1.13 Inversion Discovery

In order to assess the degree to which chromosome inversions are associated with broad-scale changes in recombination rates between zebra finch and long-tailed finch, we inferred inversion differences between zebra finch and long-tailed finch using the program DELLY (93). DELLY calls putative inversions by leveraging discordantly mapped paired-end reads and split reads to identify inversion breakpoints with respect to a reference genome. We used DELLY to predict inversions using BAM files (see *Read Processing and Alignment*) for all 19 zebra finch and all 20 long-tailed finch samples simultaneously. Minimum paired-end mapping quality was set at 20. For the 18 long chromosomes for which we had power to identify hotspots, we identified 1895 putative inversions,

which we then filtered to include only the 22 inversions that were inferred as fixed between our 19 zebra finches and 20 long-tailed finches. Representative commands used to run these analyses are available in *Representative Commands*.

1.14 Packages used for Data Visualization and Analysis

To analyze and plot these data, we used built-in Python modules, pandas v0.15 (`http://pandas.pydata.org/`), and seaborn v0.5.0 (`http://stanford.edu/~mwaskom/software/seaborn/index.html`).

1.15 Representative Commands

1.15.1 Read Mapping

Approach used to map raw reads to the reference genome.

```
bwa mem -M -t 8 taeGut1.fa 73948_R1.fastq.gz 73948_R2.fastq.gz | samtools view
    -bS - > 73948.bwamem.bam
```

1.15.2 Mark Duplicates

Approach used to identify sequence duplicates in the raw reads.

```
java -Xmx6g -jar MarkDuplicates.jar INPUT=aln-pe.73948.bwamem.bam
    OUTPUT=73948.bwamem.rmdup.bam METRICS_FILE=73948.rmdup.out
    MAX_RECORDS_IN_RAM=5000000 TMP_DIR=/tmp/
```

1.15.3 Re-align Indels

Approach used to identify indel regions and to re-align around these regions.

```
java -Xmx2g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R taeGut1.fa -I
    73948.bwamem.rmdup.bam -o 73948.bwamem.rmdup.indelsAligned

java -Xmx2g -jar GenomeAnalysisTK.jar -T IndelRealigner -R taeGut1.fa -I
    73948.bwamem.rmdup.bam -targetIntervals 73948.bwamem.rmdup.indelsAligned -o
    73948.bwamem.rmdup.realigned.bam
```

1.15.4 Fix Mate-Pair Information

Approach used to identify discordant mate-pair reads.

```
java -Xmx4g -jar FixMateInformation.jar INPUT=73948.bwamem.rmdup.realigned.bam
    OUTPUT=73948.bwamem.rmdup.realigned.mateFixed.bam
    VALIDATION_STRINGENCY=LENIENT TMP_DIR=/tmp/
```

1.15.5 Call Raw SNPs

Approach used to call the first round of variants.

```
java -Xmx4g -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R taeGut1.fasta -I
    101.mateFixed.realigned.bam -I 105.mateFixed.realigned.bam -I
    109.mateFixed.realigned.bam -I 113.mateFixed.realigned.bam -I
    117.mateFixed.realigned.bam -I 121.mateFixed.realigned.bam -I
    129.mateFixed.realigned.bam -I 133.mateFixed.realigned.bam -I
    137.mateFixed.realigned.bam -I 141.mateFixed.realigned.bam -I
    145.mateFixed.realigned.bam -I 149.mateFixed.realigned.bam -I
    153.mateFixed.realigned.bam -I 161.mateFixed.realigned.bam -I
    165.mateFixed.realigned.bam -I 173.mateFixed.realigned.bam -I
    177.mateFixed.realigned.bam -I 185.mateFixed.realigned.bam -I
    189.mateFixed.realigned.bam -L chr1 -glm BOTH -mbq 20 -hets 0.006 -out_mode
    EMIT_ALL_SITES -o chr1.raw.snps.indels.vcf
```

1.15.6 Recalibrate Raw SNPs with BQSR

Approach used to do Base Quality Score Recalibration (BQSR) of initial variants and to call new variants off recalibrated BAM files.

```
java -Xmx4g -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R taeGut1.fasta -I
    101.mateFixed.realigned.bam -knownSites all_chrs.raw.snps.indels.vcf -o
    101.recal.grp -nct 4

java -Xmx4g -jar GenomeAnalysisTK.jar -T PrintReads -R taeGut1.fasta -I
    101.mateFixed.realigned.bam -BQSR 101.recal.grp -o 101.recal.bam -nct 4

java -Xmx4g -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R taeGut1.fasta -I
    101.recal.bam -I 105.recal.bam -I 109.recal.bam -I 113.recal.bam -I
    117.recal.bam -I 121.recal.bam -I 129.recal.bam -I 133.recal.bam -I
    137.recal.bam -I 141.recal.bam -I 145.recal.bam -I 149.recal.bam -I
    153.recal.bam -I 161.recal.bam -I 165.recal.bam -I 173.recal.bam -I
    177.recal.bam -I 185.recal.bam -I 189.recal.bam -L chr1 -glm BOTH -mbq 20
    -hets 0.006 -out_mode EMIT_ALL_SITES -o chr1.post_bqsr.snps.indels.vcf -nct 4
```

1.15.7 Using Cortex for *de novo* Variant Calls

Approach used to generate *de novo* variant calls using Cortex by *de novo* assembly of each individual.

```
perl ~/cortex/releases/CORTEX_release_v1.0.5.18/scripts/calling/run_calls.pl \
   --first_kmer 31 --last_kmer 31 --kmer_step 10 \
   --fastaq_index 101.index \
   --auto_cleaning no \
   --manual_override_cleaning ../cleaning.txt \
   --bc yes --pd no \
   --outdir run_calls \
   --outvcf 101.onebyone.vcf \
   --ploidy 2 \
   --stampy_hash taeGut1 \
   --stampy_bin ~/stampy-1.0.20/stampy.py \
   --list_ref_fasta taeGut1.falist \
   --refbindir ~/finch/cortex_one_by_one/ref/ \
   --genome_size 1400000000 \
   --qthresh 5 \
   --homopol 8 \
   --mem_height 26 --mem_width 36 \
   --vcftools_dir ~/vcftools \
   --do_union yes \
   --ref CoordinatesAndInCalling \
   --workflow independent \
   --logfile 101.onebyone.log
```

1.15.8 Recalibrating Variants with VQSR

Approach used to recalibrate variants (first SNPs and then indels) with Variant Quality Score Recalibration (VQSR) using the trusted variant set of the intersection of Cortex and initial GATK variant calls.

```
java -Xmx20g -jar GenomeAnalysisTK.jar -T VariantRecalibrator -R taeGut1.fa
    -input all_chrs.post_bqsr.snps.indels.vcf.gz
    -resource:GATK_cortex_intersection,known=true,training=true,truth=true,prior=10.0
    GATK_cortex_intersection.all_chrs.snps.indels.vcf.gz -an QD -an
    HaplotypeScore -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an DP
    -recalFile all_chrs.snps.recal -tranchesFile all_chrs.snps.tranches
    -rscriptFile all_chrs.VQSR.snps.plots.R -mode SNP -nt 8

java -Xmx20g -jar GenomeAnalysisTK.jar -T ApplyRecalibration -R taeGut1.fa -input
    all_chrs.post_bqsr.snps.indels.vcf.gz --ts_filter_level 99.0 -recalFile
```

```
    all_chrs.snps.recal -tranchesFile all_chrs.snps.tranches -o
    all_chrs.vqsr.snps.indels.vcf.gz -mode SNP -nt 8

java -Xmx20g -jar GenomeAnalysisTK.jar -T VariantRecalibrator -R taeGut1.fa
    -input all_chrs.vqsr.snps.indels.vcf.gz
    -resource:GATK_cortex_intersection,known=true,training=true,truth=true,prior=10.0
    GATK_cortex_intersection.all_chrs.snps.indels.vcf.gz -an QD -an MQRankSum -an
    ReadPosRankSum -an FS -an DP -recalFile all_chrs.indels.recal -tranchesFile
    all_chrs.indels.tranches -rscriptFile all_chrs.VQSR.indels.plots.R -mode
    INDEL -nt 8

java -Xmx20g -jar GenomeAnalysisTK.jar -T ApplyRecalibration -R taeGut1.fa -input
    all_chrs.vqsr.snps.indels.vcf.gz --ts_filter_level 99.0 -recalFile
    all_chrs.indels.recal -tranchesFile all_chrs.indels.tranches -o
    /all_chrs.vqsr2.snps.indels.vcf.gz -mode INDEL -nt 8
```

1.15.9 Identify Mendelian Errors

Approach used to identify Mendelian errors in the family of zebra finches.

```
plink --noweb --file all_zf.chr1 --mendel --out all_zf.me.chr1
```

1.15.10 Get Species Genomes

Approach used to get variant calls for the medium ground finch (*Geospiza fortis*), large ground finch (*Geospiza magnistrosis*), and collared flycatcher (*Ficedula albicollis*).

```
bowtie2 -D 20 -R 3 -N 1 -L 20 -i S,1,0.50 -p 8 --local -x taeGut1.fasta -1
    Ficedula_albicollis_1.fastq.gz -2 Ficedula_albicollis_2.fastq.gz -S
    Ficedula_albicollis.sam

samtools mpileup -I -uf taeGut1.fasta Ficedula_albicollis.bam | bcftools view -t
    0.1 -I -c - >Ficedula_albicollis.vcf
```

1.15.11 Family Phasing

Approach used to phase the family of zebra finches.

```
hapi-mr -h -d output_files/ chr1.hapi.list chr1.hapi.sites chr1.hapi.gen
```

1.15.12 Population Phasing

Approach used to phase each species. This command was used for zebra finch, for which we had reference haplotypes from the family phasing.

```
extractPIRs --bam list_of_bam_files.txt --vcf chr1.vcf.gz --out chr1_PIRlist

shapeit -assemble --input-vcf chr1.vcf.gz --input-pir chr1_PIRlist -O
    chr1_haplotypes -L chr1_haplotypes.log --window 0.5 --thread 8 --rho 0.0008
    --output-graph chr1_haplotypes.graph -R chr1.hap.gz chr1.legend.gz chr1.sample
```

1.15.13 Infer Recombination Maps

Approach used to infer recombination rates. The 'prior-rate' flag was set using chromosome and species-specific estimates from preliminary runs of LDhelmet.

```
ldhelmet rjmcmc --num_threads 12 -o recombination_map -n 1000000 --burn_in 100000
    -b 5 -s haplotypes.fasta -l species_likelihood_table -p species_pade -a
    ancestral_alleles.txt -m mutation_matrix.txt -w 50 --max_lk_end 100
    --prior_rate 0.05
```

1.15.14 Motif Discovery for Hotspots

Approach used to identify motifs for hotspots. The first step allows for discriminatory motif analysis between hotspots and coldspots.

```
~/bin/psp-gen -pos hotspot.subset.fa -neg coldspot.subset.fa -revcomp > motifs.psp

~/bin/meme hotspot.subset.fa -psp motifs.psp -oc out_dir/ -revcomp -dna -nmotifs
    10 -minw 5 -maxw 20 -maxsize 1050000
```

1.15.15 Identify Inversions

Approach used to identify inversions in each individual with respect to the reference genome.

```
delly -t INV -x Delly_exclude.list -q 20 -o EstrildidFinch_INVY.vcf -g taeGut1.fa
    <ALL_BAM_FILES>
```

2 Supplementary Figures

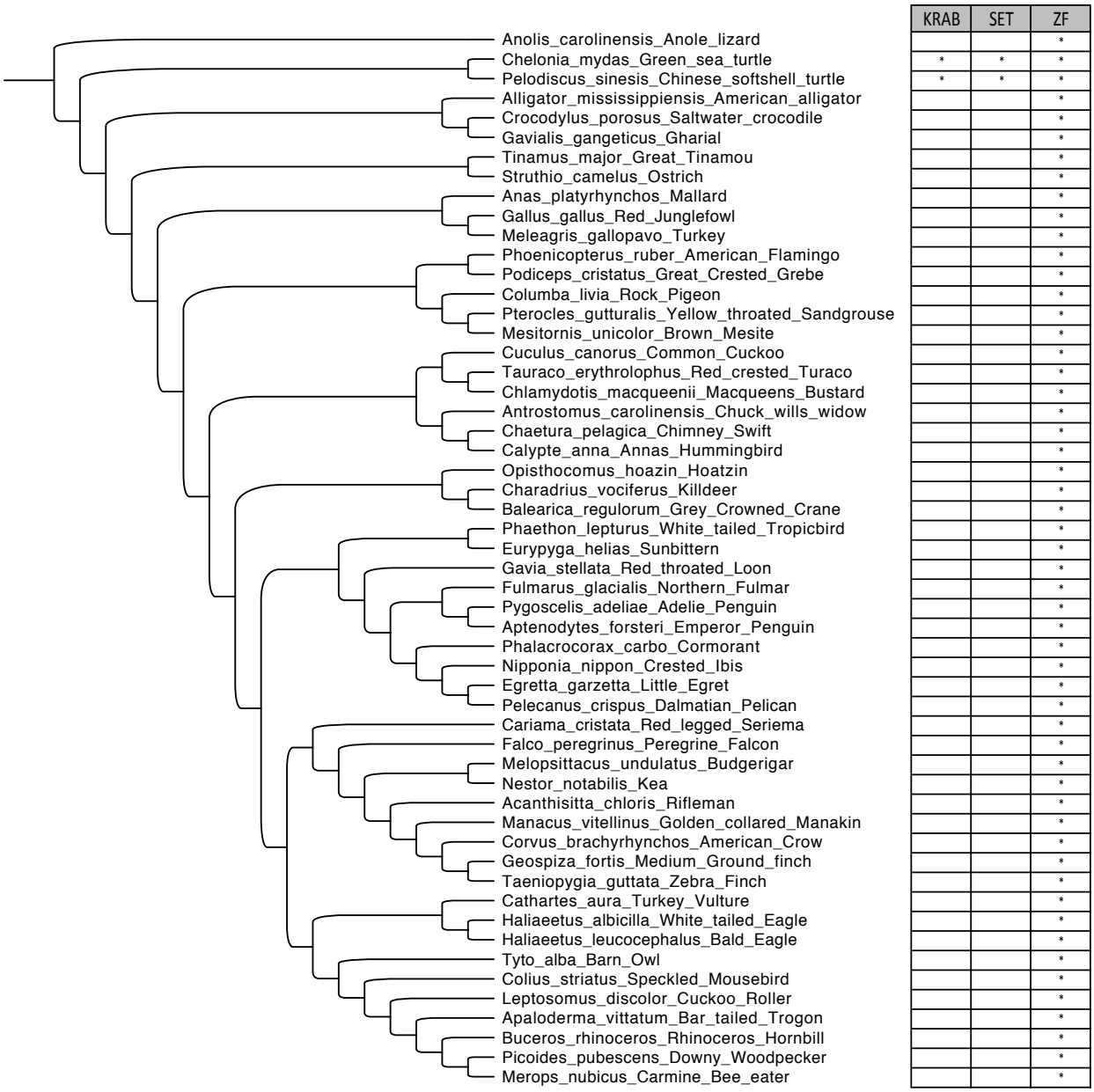

Figure S1: Species for which the presence of PRDM9 was tested, which include 48 birds, two turtles, one lizard, and three crocodilians. Presence or absence of the three domains of PRDM9, as determined by BLAST searches, are shown; matches in species with only a zinc finger domain (ZF) are unlikely to correspond to PRDM9. Phylogeny taken from (53).

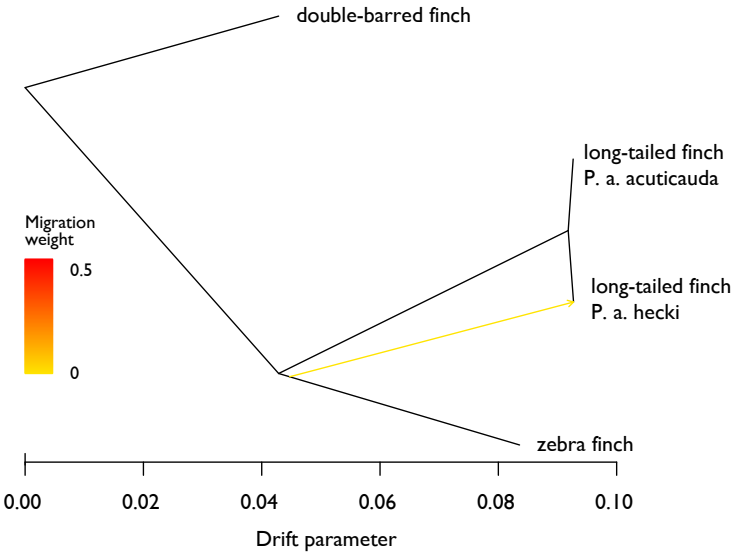

Figure S2: TreeMix results for zebra finch (*Taeniopygia guttata*), the two subspecies of long-tailed finch (*Poephila acuticauda acuticauda* and *P. a. hecki*), and double-barred finch (*T. bichenovii*). Under the four-population test, the migration inferred is non-significant (z-statistic=0.98; p=0.32; see *Species Tree Inference and TreeMix*).

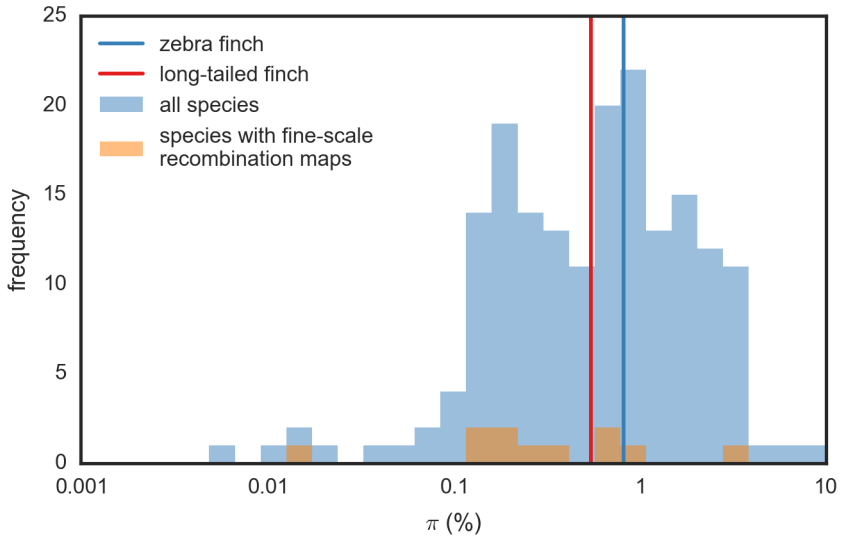

Figure S3: Nucleotide diversity (π) across a wide range of species (as reported in (20)) and for which previous studies have inferred fine-scale recombination rates. Also shown is π for the two species for which we infer fine-scale recombination maps in this work, zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*).

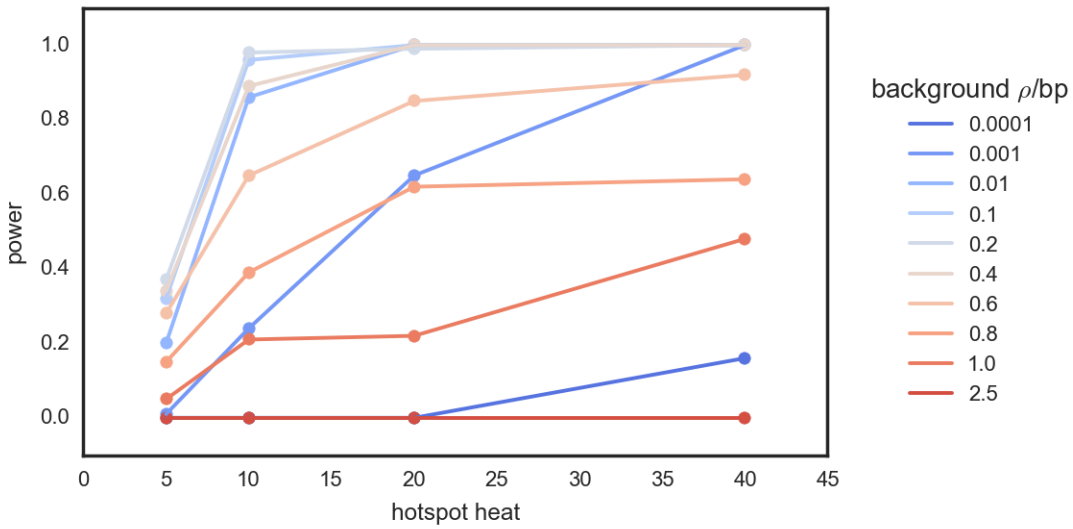

Figure S4: Power to identify hotspots for different background ρ. Simulations were run for a range of background ρ with hotspots of varying heats that reflect the range seen in zebra finch (*Taeniopygia guttata*). Rates were estimated using LDhelmet under block penalty 5; hotspots were inferred; and power was calculated. For each parameter set, 100 hotspots were simulated. These results indicate that we have limited power to detect hotspots at high ($\rho > 0.8$/bp) and low ($\rho \leq 0.0001$/bp) values of background ρ.

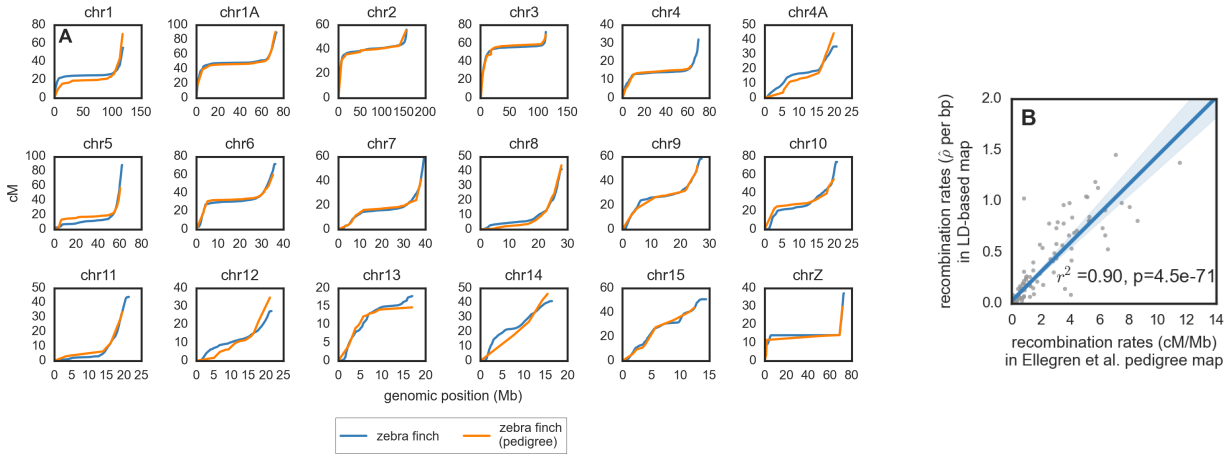

Figure S5: (A) Relationship between genetic and physical maps for zebra finch (*Taeniopygia guttata*) as inferred in this study and pedigree data from (21). See *Comparison to an Existing Genetic Map* for details on the conversion of $\hat{\rho}$ to cM/Mb. (B) Correlation between recombination rates as inferred in (21) in cM/Mb and our map in $\hat{\rho}$ per bp across 5 Mb windows, along with the fit of a linear regression.

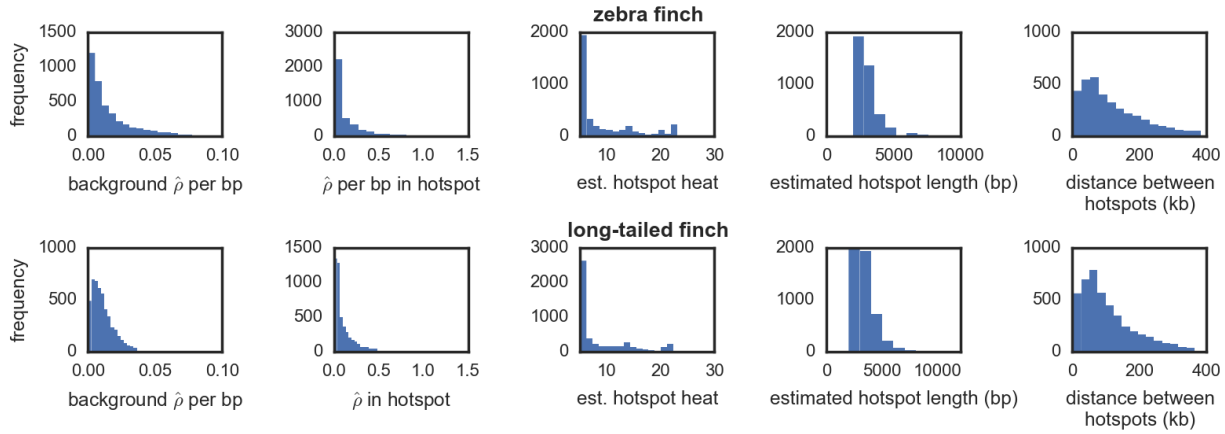

Figure S6: Characteristics of hotspots inferred in (top) zebra finch (*Taeniopygia guttata*) and (bottom) long-tailed finch (*Poephila acuticuada*), including the background $\hat{\rho}$ of the hotspots, $\hat{\rho}$ within hotspots, the heat of the hotspots, the length of hotspots, and the distance between adjacent hotspots on the same chromosome. Hotspots were defined as being $\geq$2 kb in length and with $\hat{\rho}$ at least $\geq$5 the background rate, which was inferred for 40 kb of sequence from the hotspots in each direction. Estimated hotspot heat was defined with respect to the same background $\hat{\rho}$. Distributions plotted here have long tails; to ease the visualization of the data, we exclude the top 5% of each distribution.

We note that hotspot density appears to be lower in finches compared to humans (86) and hotspot intensity in finches appears to be lower than in humans and apes (86; 9). The lower density of hotspots in the finches compared to humans is consistent with simulations that indicate decreased power to detect hotspots when the background population recombination rate is higher (Figs. S4, S8). Further, simulations suggest the two-fold lower average hotspot intensity relative to apes could in part be due to a downward bias in estimates (Fig. S30). However, a non-mutually exclusive possibility for both patterns is that the finches' fine-scale recombination landscape is truly less punctate than in apes.

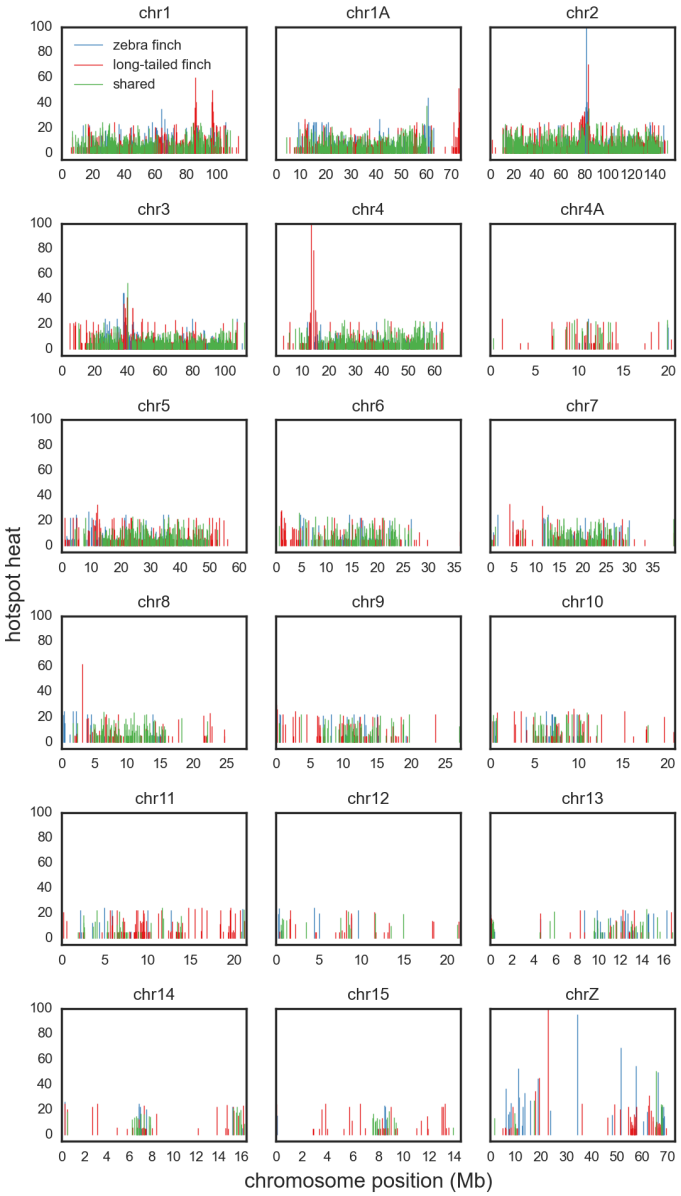

Figure S7: Location of hotspots across chromosomes for hotspots detected only in zebra finch (*Taeniopygia guttata*), only in long-tailed finch (*Poephila acuticauda*), and those detected as shared between the species. The height of the line indicates the hotspot heat.

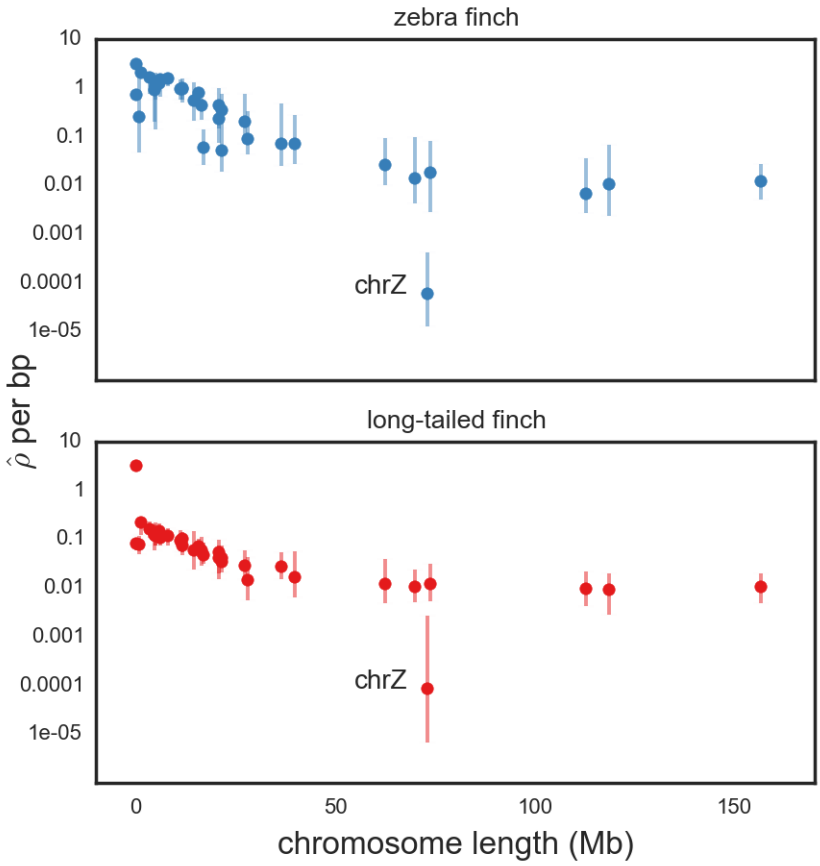

Figure S8: Median estimated recombination rate ($\hat{\rho}$) for a given chromosome as a function of chromosome length in zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). Error bars show upper and lower quartiles of $\hat{\rho}$. To enable comparisons across chromosomes, $\hat{\rho}$ for chromosome Z is shown doubled, because $\rho_{chrZ}=2N_e c$ and $\rho_{autosomes}=4N_e c$ under a simple neutral model, where N_e is x and c is y. Rate estimates for chromosome Z should be taken with caution (see *Variant Quality*).

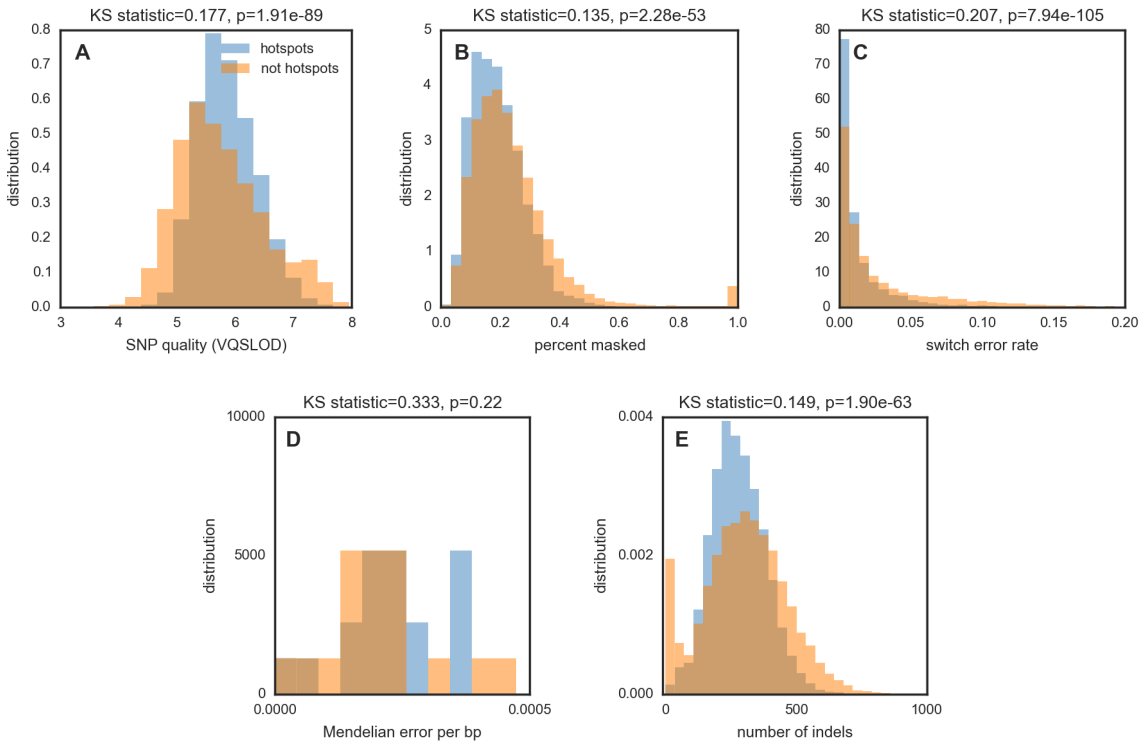

Figure S9: Comparisons of data quality in and near hotspots compared to non-hotspot sequence in zebra finch (*Taeniopygia guttata*): (A) variant quality (as reflected in SNP VQSLOD score reported by GATK) in 50 kb windows surrounding hotspots versus other 50 kb windows, (B) percentage of sequence masked in 50 kb windows surrounding hotspots versus other 50 kb windows, (C) rate of switch error (as estimated for the parents of the domesticated zebra finch family) in 50 kb windows surrounding hotspots versus other 50 kb windows, (D) Mendelian error rate (as estimated for the domesticated zebra finch family) in 50 kb windows surrounding hotspots versus other 50 kb windows, and (E) number of indels in 50 kb windows surrounding hotspots versus other 50 kb windows. Reported above each graph is the Kolmogorov-Smirnov test for equality of distributions and its significance; importantly, although the distributions of values differs significantly in (A), (B), (C) and (E), hotspots have higher data quality than non-hotspot sequence in each case.

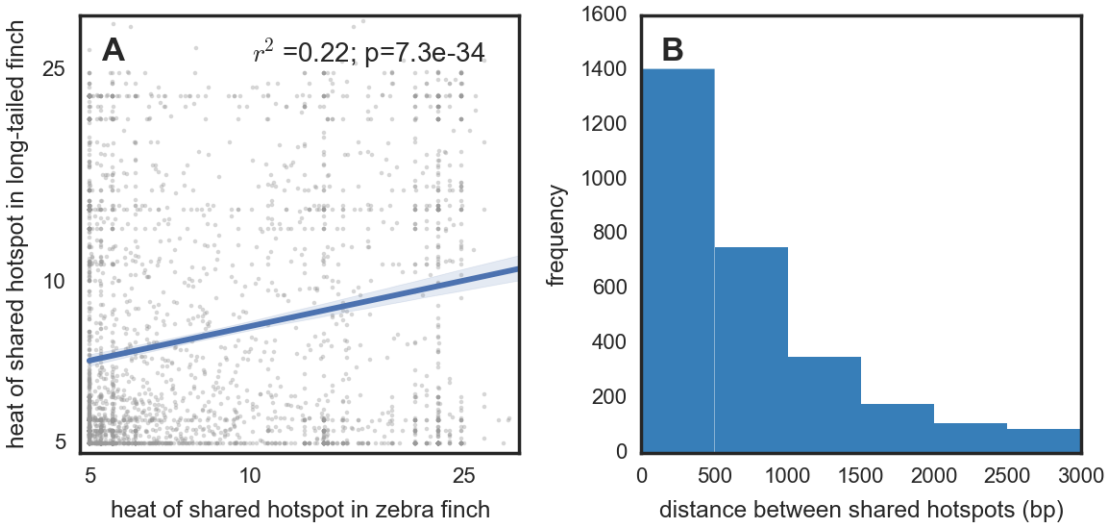

Figure S10: Descriptive plots for hotspots detected as shared between zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). Shared hotspots are defined as those whose midpoints occur within 3 kb of each other. (A) Correlation in hotspot heats between the two species. Best-fit linear regression shown with 95% confidence interval and its fit. (B) Distance between midpoints of shared hotspots in the two species.

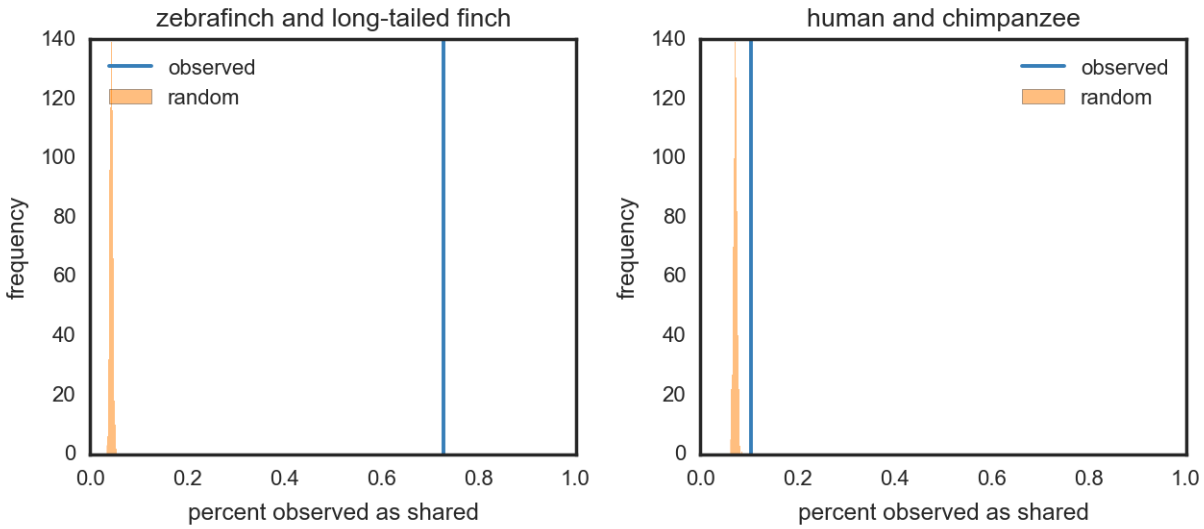

Figure S11: Null expectations and observed values for hotspot sharing between zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*) and between human and chimpanzee, defining hotspots as shared if their midpoints are within 3 kb of each other. Percent shared is calculated using the number of hotspots in the species with fewer hotspots as the denominator. Observed sharing between human and chimpanzee is only marginally above naive null expectations, whereas observed sharing between zebra finch and long-tailed finch is many-fold higher than expected under the null.

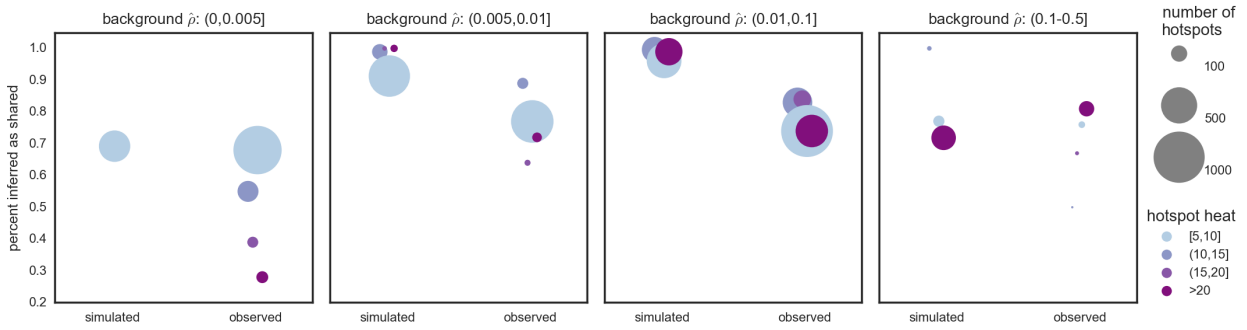

Figure S12: Percentage of hotspots that are inferred as shared when all hotspots are simulated as having shared heats and locations. For each species, data were simulated under θ values reflective of the two focal species in this study, zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). Because zebra finch has approximately two-fold greater ρ than long-tailed finch, one population was simulated with double the ρ of the other species. Shown are $\hat{\rho}$ for the species with the higher rate, which were simulated under ρ ranging from 0.001, 0.002, 0.01, 0.1, 0.5, 0.8. During simulation of the sequence, hotspots were placed in the same locations in the two species. Recombination rates are inferred using LDhelmet under block penalty 5; putative hotspots were inferred in both species; and hotspots were validated using sequenceLDhot. After binning hotspots based on their estimated background $\hat{\rho}$ and their estimated heat, the percent shared (i.e., those hotspots for which the distance between midpoints is $\leq$ 3 kb) was calculated. We note that, although we ran simulations for all bins, we report estimated values, not simulated values, so some bins have no data.

Also shown are the observed levels of hotspot sharing between zebra finch and long-tailed finch; hotspot sharing between the two finches is lower than expected compared to simulated results.

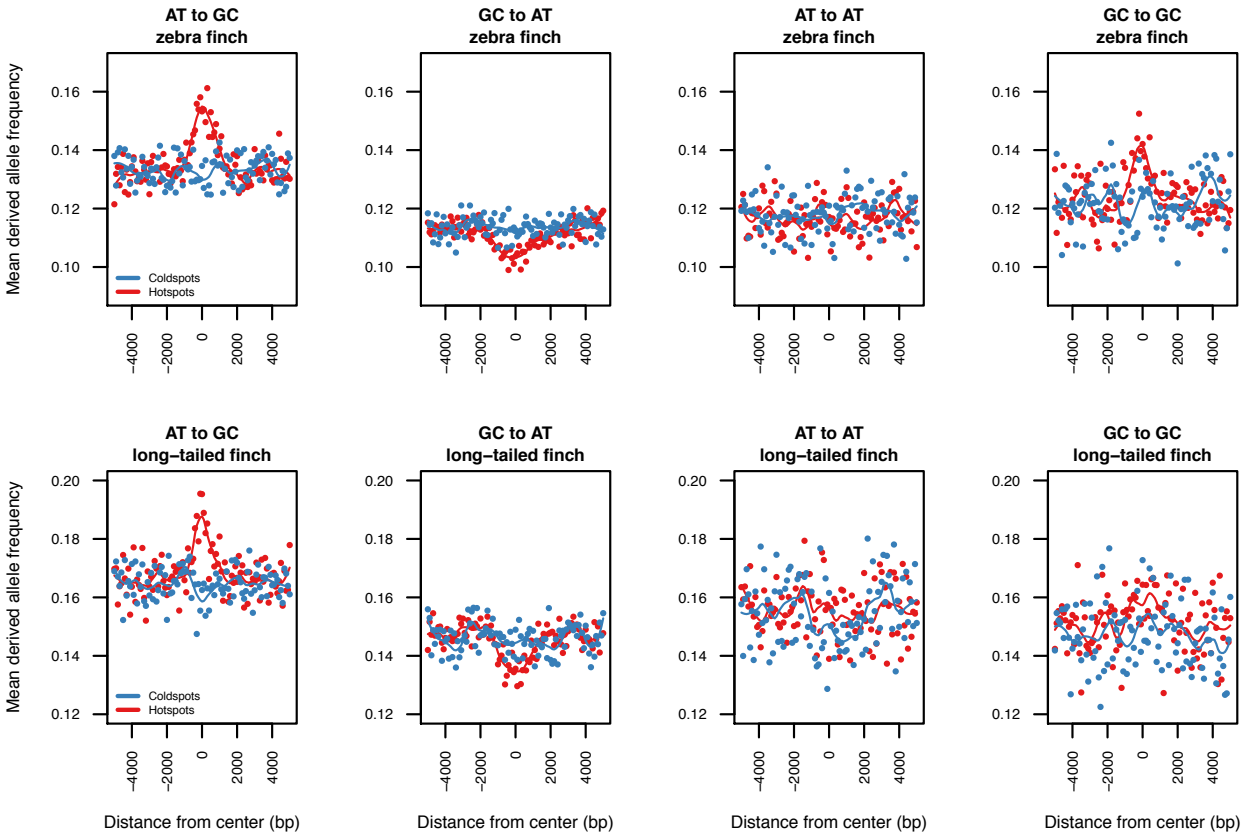

Figure S13: Mean derived allele frequency for SNPs of different mutation types around shared hotspots and matched coldspots in zebra finch (*Taeniopygia guttata*, top) and long-tailed finch (*Poephila acuticauda*, bottom). Variants at all potential CpG sites (where either allele creates a CpG in the ancestral sequence) were excluded because they are more liable to ancestral misidentification. In addition to AT to GC mutations, there is some indication of a trend towards higher derived allele frequency for GC to GC mutations at hotspots; although we do not have an explanation for this observation, we also note that an excess of GC to GC SNPs has been reported at double-strand break hotspots in humans (94).

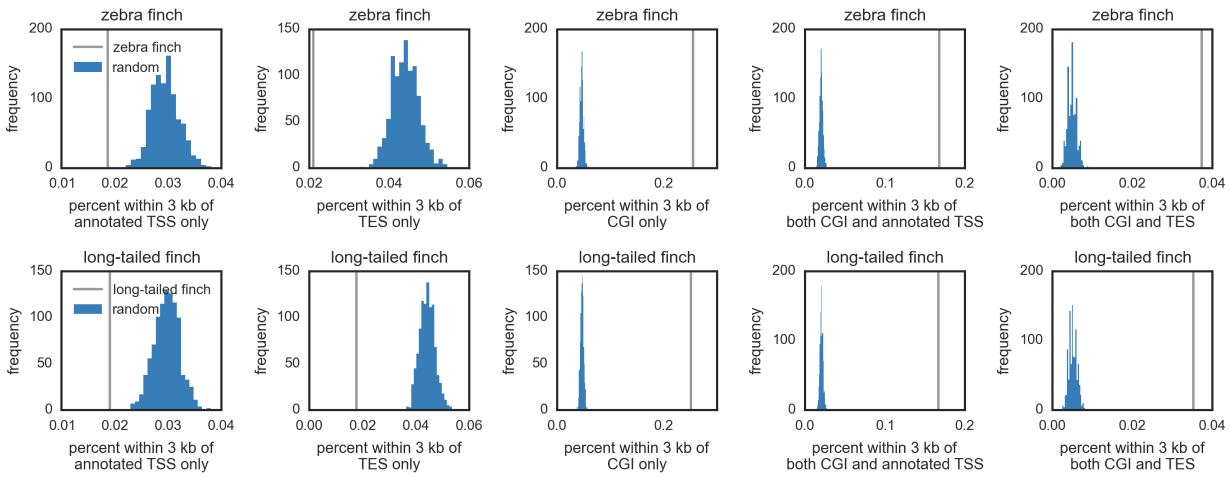

Figure S14: Location of hotspots in zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*) with respect to annotated transcription start sites (TSSs), transcription end sites (TESs), and CpG islands (CGI) plotted by comparison to locations for 1000 randomly drawn spots. Each set of random spots was drawn proportionally to the distribution of inferred hotspots across chromosomes. Distances of both inferred hotspots and random hotspots to TSS and CGI features were summarized by percent of spots within 3 kb of a feature. Vertical lines indicate distances found in this study.

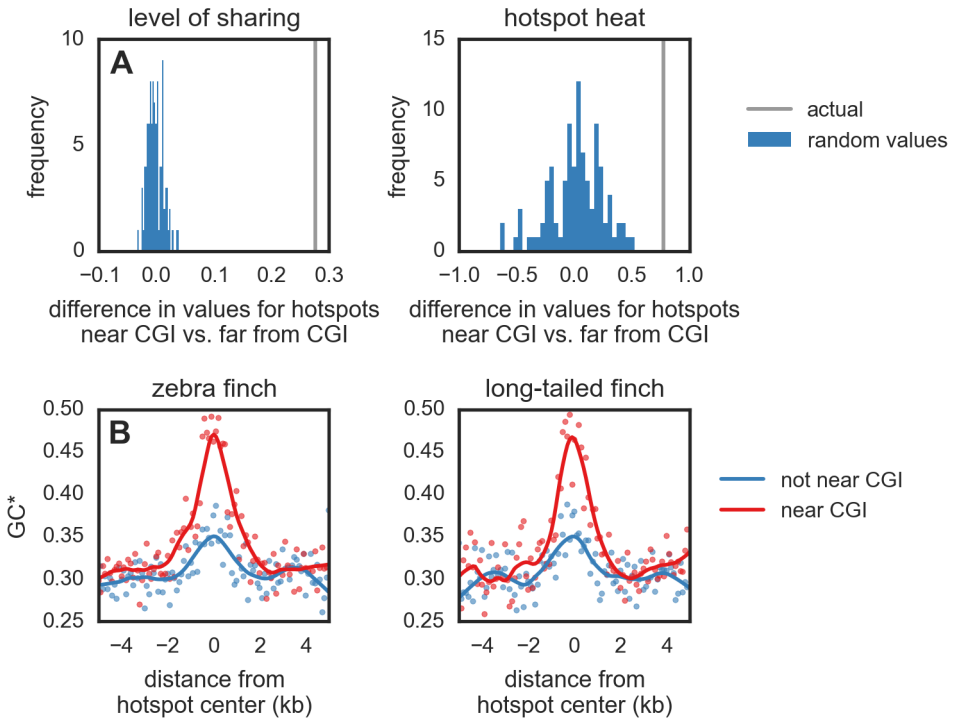

Figure S15: Properties of hotspots within 3 kb of CpG islands (CGIs) compared to those distant from CGIs in zebra finch (*Taeniopygia guttata*). (A) Percent of hotspots inferred as shared (i.e., having midpoints within 3 kb of each other) between zebra finch and long-tailed finch (*Poephila acuticauda*) and hotspot heat as measured by sequenceLDhot. Each plot shows the difference in mean (gray line) for this metric between hotspots near CGIs (for zebra finch, n=1,709; long-tailed finch, n=2,122) and hotspots that are not (for zebra finch, n=2,240; long-tailed finch, n=2,811). The histogram (shown in blue) indicates the difference in means calculated for 100 bootstrap samples in which hotspots were randomly assigned to being near CGIs or not, following their empirical proportions. (B) Expected equilibrium GC content (GC*) around hotspots in zebra finch and long-tailed finch near CGIs and those far from CGIs. Points represent GC* estimated from the lineage-specific substitutions aggregated in 100 bp bins from the center of all hotspots; LOESS curves are shown for a span of 0.2. The orientation of hotspots is with respect to the genomic sequence.

These results indicate that hotspots near CGIs are more likely to be shared, are hotter, and exhibit greater GC* than hotspots far from CGIs.

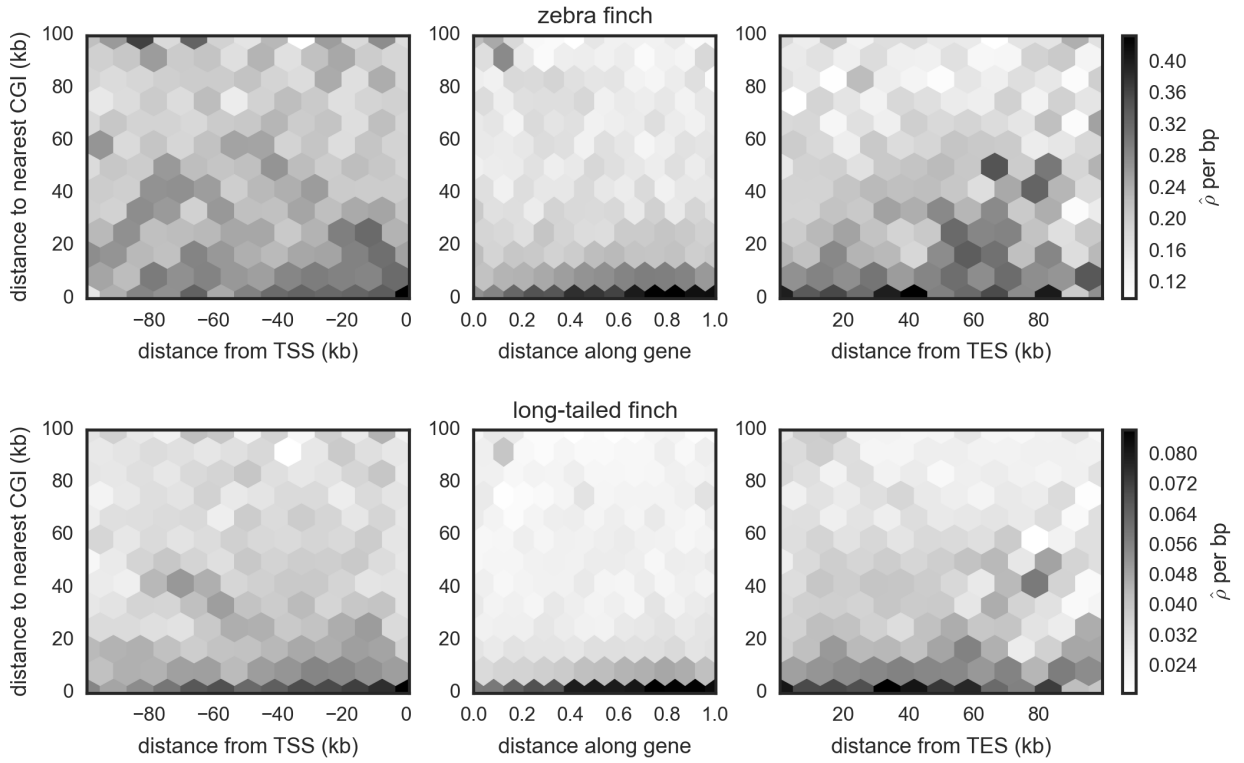

Figure S16: Recombination rate ($\hat{\rho}$ per bp) measured with respect to position proximate to genes' annotated transcription start sites (TSSs) and transcription end sites (TES) and the nearest CpG island (CGI) for zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). Distance along a gene reflects the proportional location of a given position with respect to total gene length, and this plot includes exons and introns. Rates around the TSS and TES reflect the 5′→3′ orientation of genes. These results suggest that distance to CGI affects $\hat{\rho}$ more than does distance to annotated TSS and TES.

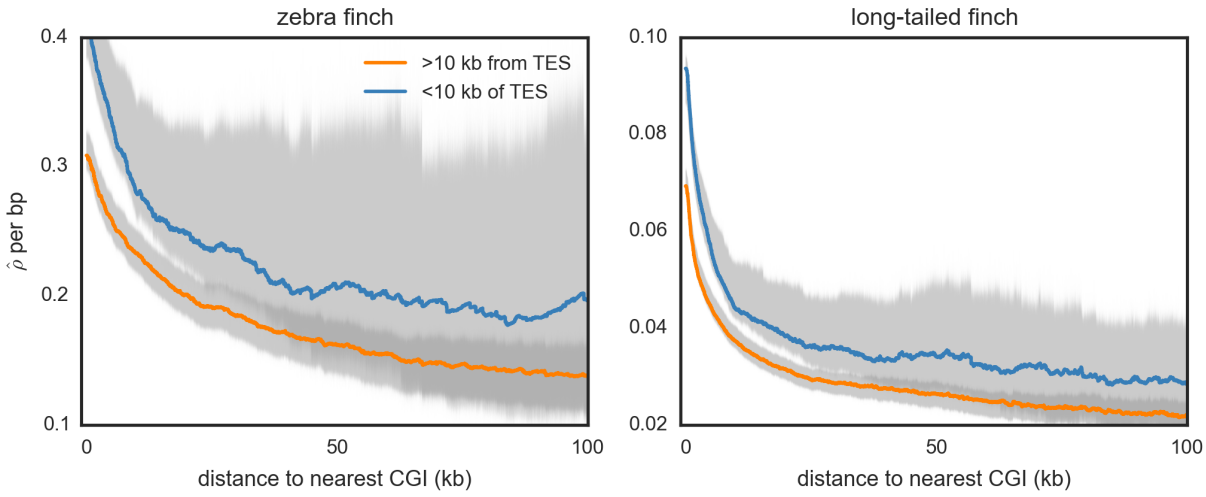

Figure S17: For zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*), estimated recombination rates ($\hat{\rho}$ per bp) around a CpG island (CGI), conditional on whether they are within 10 kb of a transcription end site (TES). Uncertainty in rate estimates (shown in gray) was estimated by drawing 100 bootstrap samples and recalculating means.

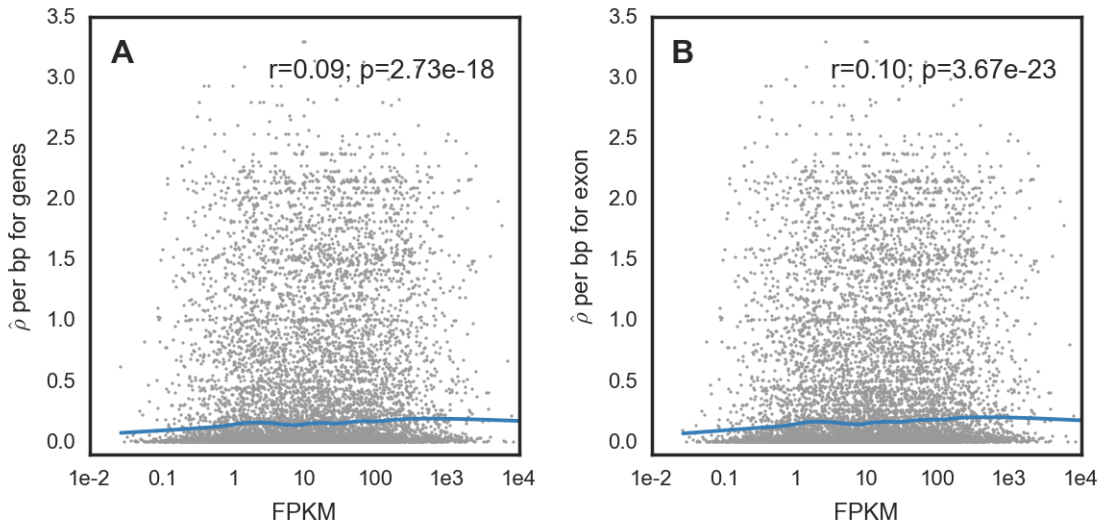

Figure S18: The relationship between gene expression and estimated recombination rates ($\hat{\rho}$). Gene expression is measured as fragments per kilobase of exon per million reads mapped (FPKM) for 9,281 genes averaged across six zebra finch (*Taeniopygia guttata*) testes RNAseq samples sequenced with Illumina (see *Analysis of Gene expression*). (A) $\hat{\rho}$ were estimated per gene, from transcription start site to end site. (B) $\hat{\rho}$ measured for the first exon. For both (A) and (B), we report Spearman's rank correlation between FPKM and $\hat{\rho}$ and a LOESS with span 0.2.

Because distance to CpG island (CGI) is correlated with recombination rate (Fig. 5), and recombination rate is correlated with gene expression levels, we further confirm that recombination rate and CGI remain correlated when controlling for gene expression levels, finding a Spearman's partial $r = -0.1; P = 4.32 \times 10^{-27}$.

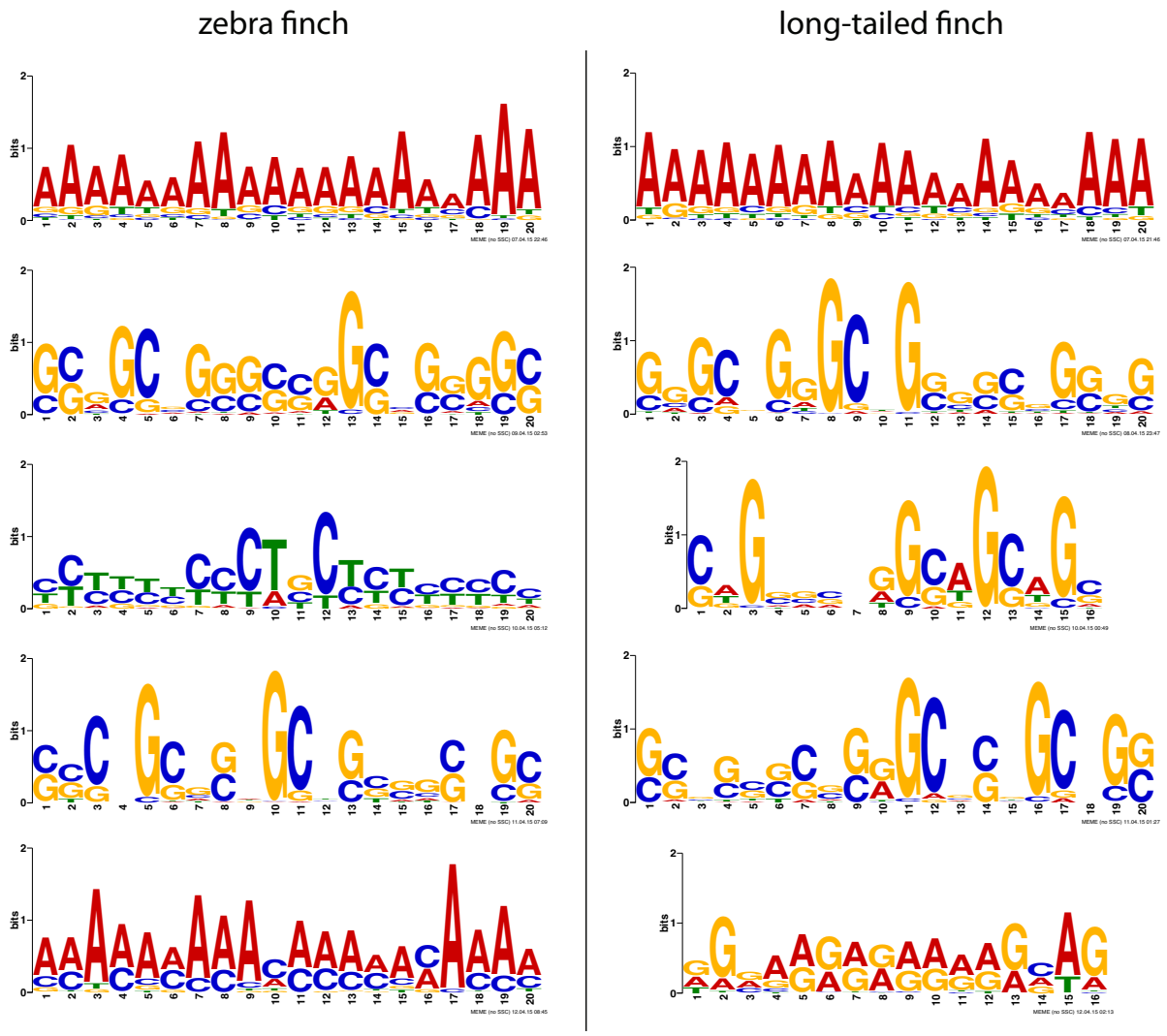

Figure S19: Top motifs discovered by MEME as being significantly associated with hotspots in zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). MEME was run using a randomly selected set of 1000 hotspots for each species in discriminative mode using coldspot sequences. This analysis was replicated across five runs; motifs identified were similar across runs. The top two motifs were the same in the other run, the other motifs were identified in the top ten motifs in all runs. Shown are the results from the first run.

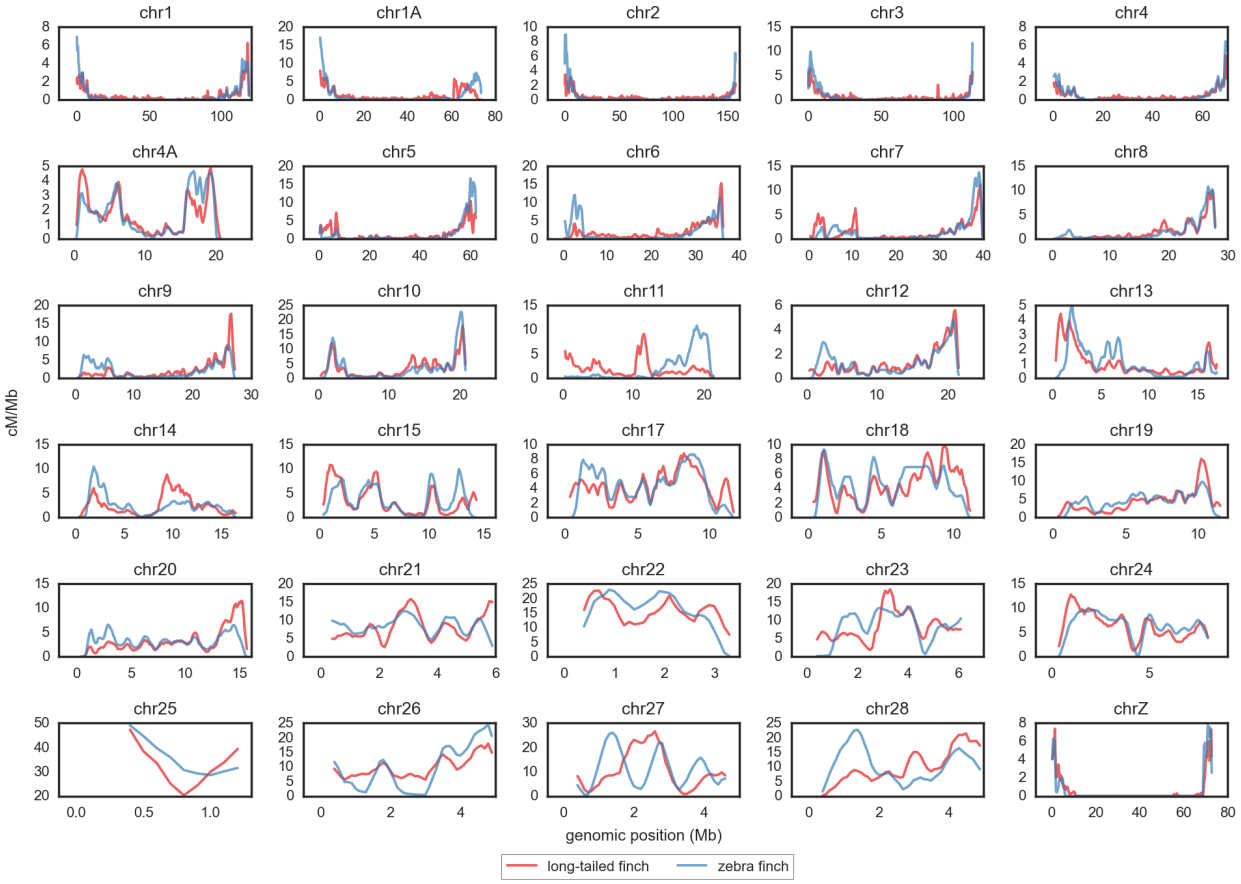

Figure S20: Estimated recombination rate (cM/Mb) for zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*) shown as rolling means calculated across ten 100 kb windows. Chromosomes 16, LG2, LG5, and LGE22 are not shown because LDhelmet runs for these chromosomes failed, likely because these chromosomes are very short (<1 Mb). See *Comparison to an Existing Genetic Map* for details on how $\hat{\rho}$ for zebra finch and long-tailed finch was converted to cM/Mb. Rate estimates for chromosome Z should be taken with caution (see *Variant Quality*).

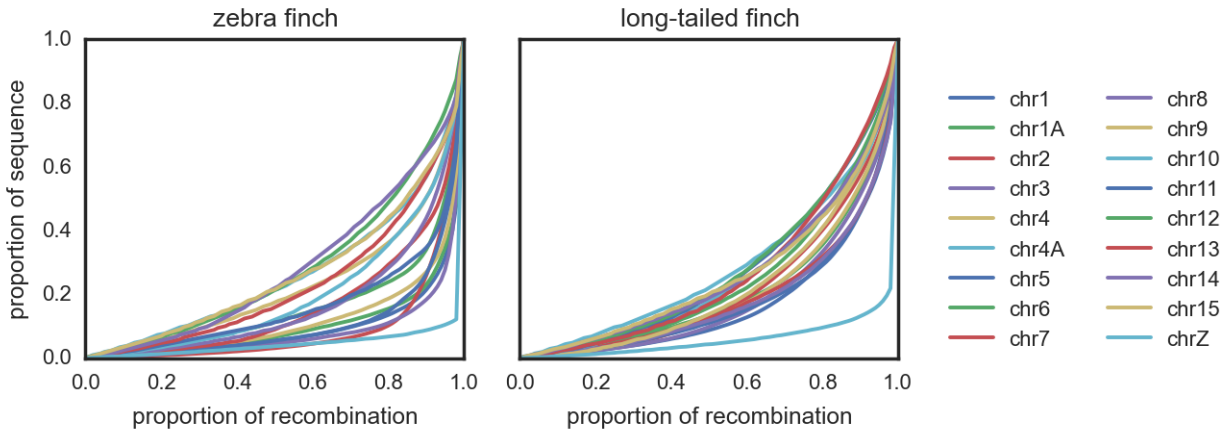

Figure S21: Proportion of recombination in a given proportion of sequence in zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). We show only the 18 chromosomes for which we expect to have power to detect recombination hotspots (see *Power to Detect Hotspots*). Rate estimates for chromosome Z should be taken with caution (see *Variant Quality*).

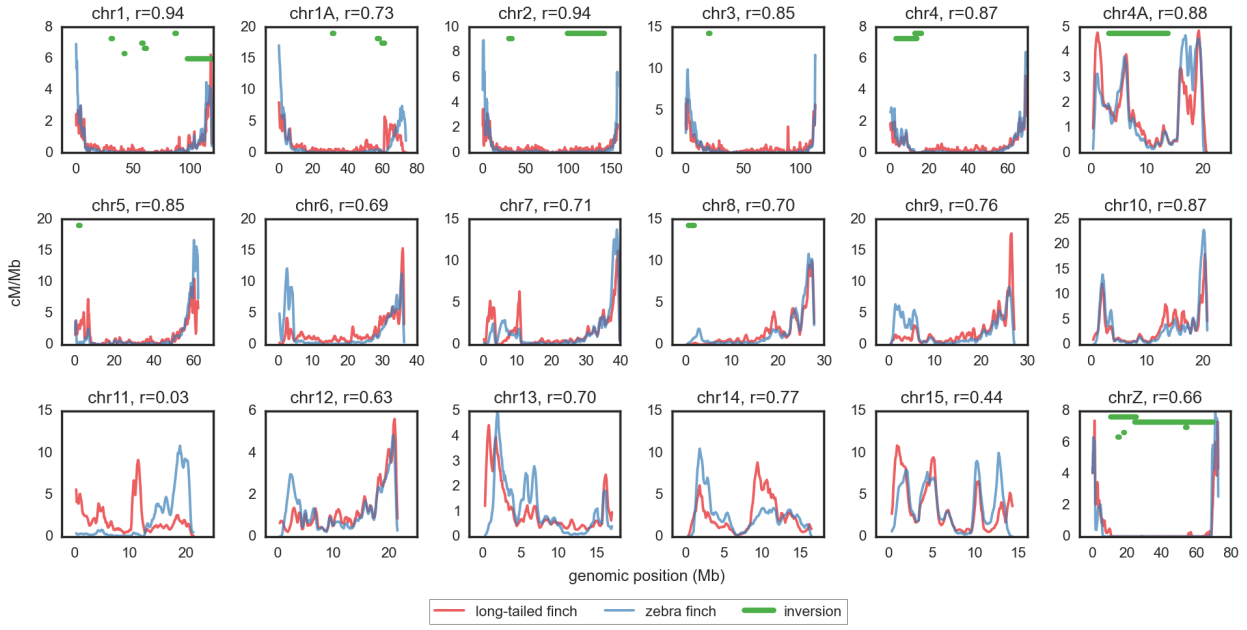

Figure S22: Estimated recombination rate (cM/Mb) for zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*) shown as rolling means calculated across ten 100 kb windows for the 18 chromosomes for which we expect to have power to detect recombination hotspots (see *Power to Detect Hotspots*). See *Comparison to an Existing Genetic Map* for details on how $\hat{\rho}$ for zebra finch and long-tailed finch were converted to cM/Mb. Rate estimates for chromosome Z should be taken with caution (see *Variant Quality*).

Spearman's correlation of zebra finch and long-tailed finch rates measured for 1 Mb windows is reported for each chromosome. Out of 937 1 Mb windows on the 18 longest chromosomes, 278, 128, 23, and 16 windows have recombination rates that differ 5-fold, 10-fold, 50-fold, and 100-fold respectively between the two species.

Fixed inversions between zebra finch and long-tailed finches detected using the program DELLY are shown. These data show that regions with big differences in recombination rate between zebra finch and long-tailed finch are not significantly enriched for fixed inversions. In fact, Kolmogorov-Smirnov test for equality of distributions indicates that regions with fixed inversions tend to have differences in recombination rate that are relatively less severe than regions without fixed inversions (KS statistic=0.12; $p = 4.28 \cdot 10^{-12}$).

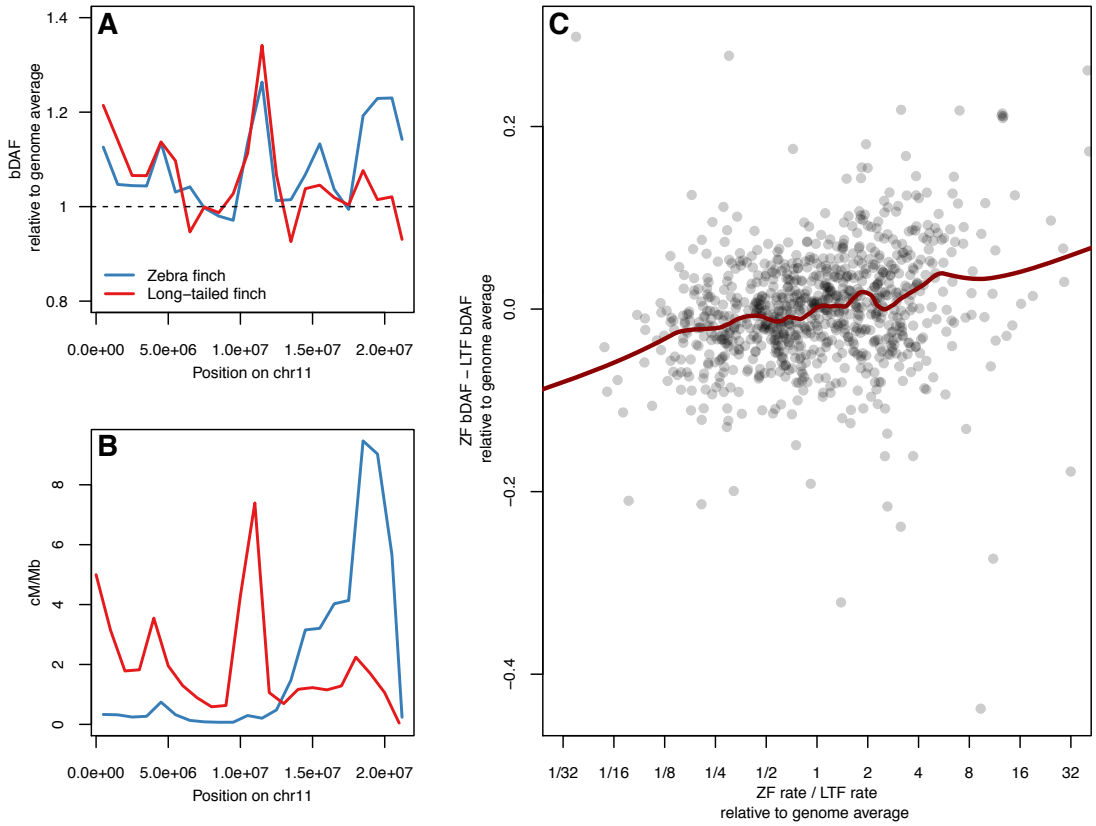

Figure S23: Impact of biased gene conversion on the derived allele frequency (DAF) in regions where broad-scale recombination rates differ between zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). To quantify the effect of GC-biased gene conversion (gBGC) on allele frequencies, we calculated the mean DAF at AT→GC SNPs. Because the level of polymorphism varies across the genome, we then divided this by the mean DAF at AT→AT SNPs, which is unaffected by gBGC, and call this quantity bDAF. (A) bDAF in 1 Mb windows for zebra finch and long-tailed finch, normalized by the mean value across the genome for each species. (B) Recombination rate in 1 Mb windows across chr11 for zebra finch and long-tailed finch, where several regions show recombination rate differences between species. We see that the region in the middle of the chromosome, where long-tailed finch has a higher recombination rate than zebra finch, shows increased bDAF in both species suggesting that the recombination rate is similarly high in both, in contrast to the rate estimates. The region at the end of the chromosome, on the other hand, only shows an increase in bDAF in zebra finch, consistent with a true difference in recombination rates between species. (C) Difference in recombination rate and difference in bDAF between species across 1Mb windows on the 17 largest autosomes. Relative rate of recombination, on the x axis, is measured as the ratio of the rate in zebra finch to the rate in long-tailed finch, normalized by the average ratio across all windows and is shown on a log2 scale. The difference in bDAF between the two species is also normalized by the average ratio across the genome. There is a trend for higher bDAF in the species in which recombination rates are estimated to be higher, consistent with true differences in recombination rate between the species at this scale. A LOESS curve is shown for a span of 0.2 and the x-axis has been truncated, excluding extreme values.

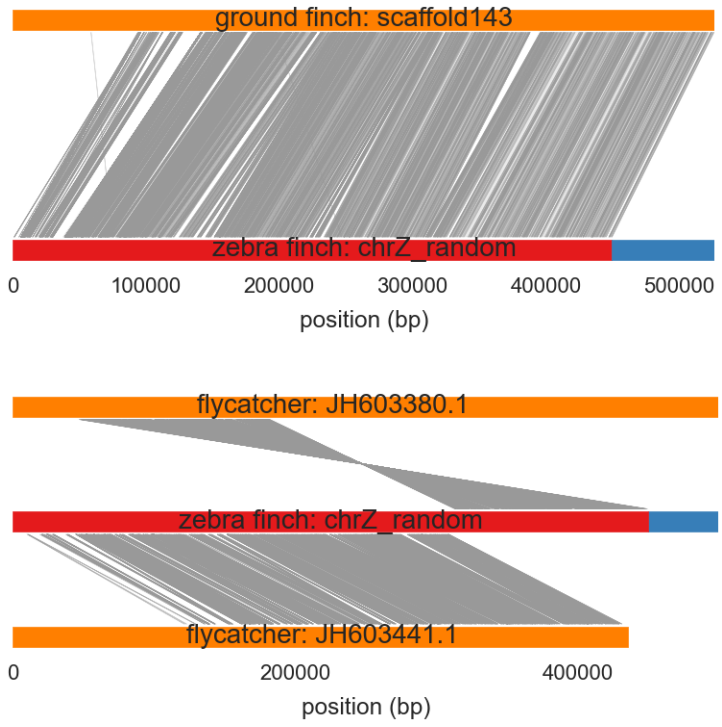

Figure S24: Alignments of the pseudoautosomal region (PAR; shown in red) in zebra finch (*Taeniopygia guttata*) with the PAR identified in the medium ground finch (*Geospiza fortis*; (63)) and in the collared flycatcher (*Ficedula albicollis*; (64)). (64) reported that three contigs comprise the PAR in the collared flycatcher; however, we were unable to find one contig (AGTO1003702.1) in the genome assembly. The remaining two contigs are shown here. Sequences were aligned using LASTZ with default settings.

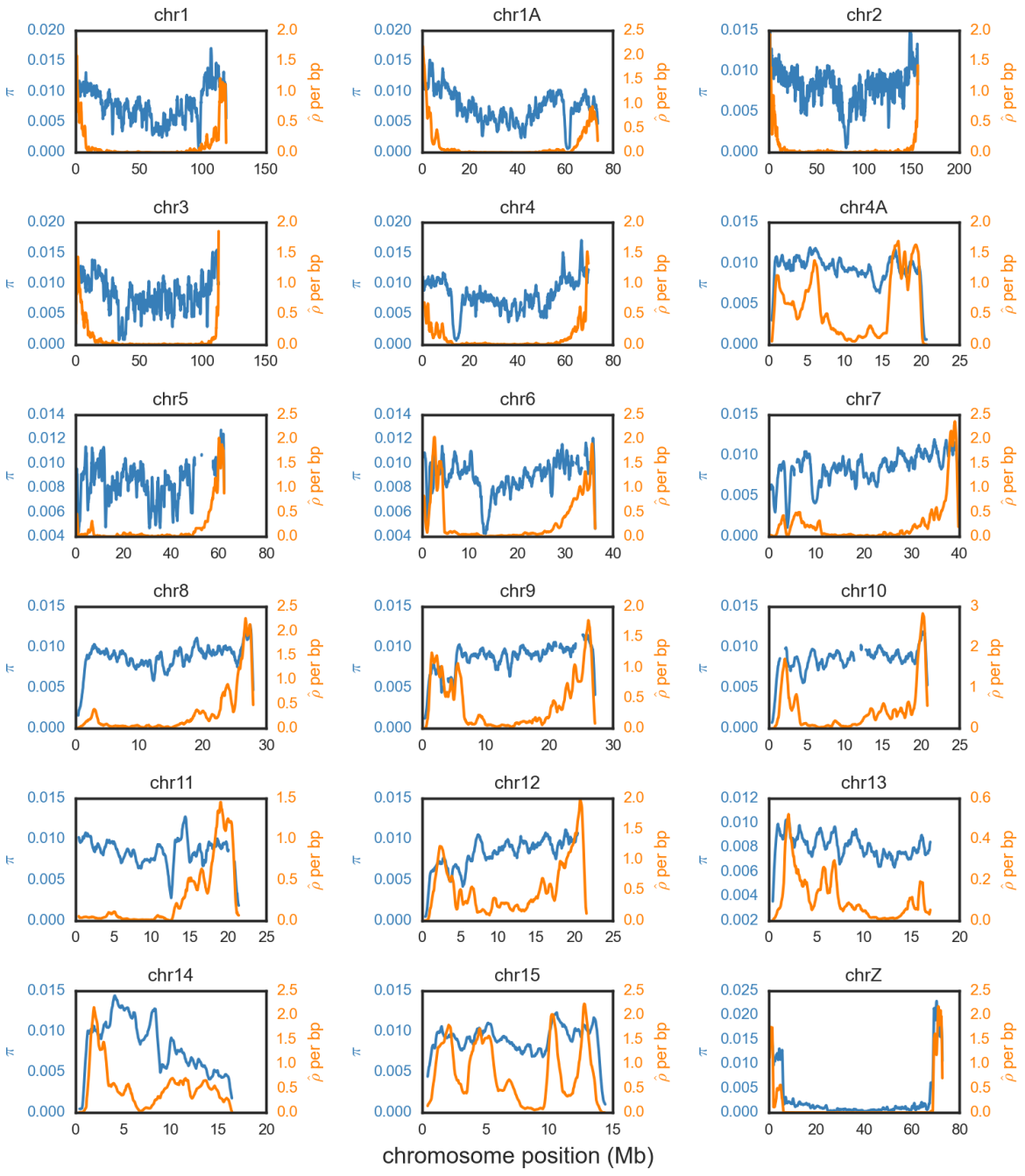

Figure S25: Estimated nucleotide diversity (π) and recombination rate ($\hat{\rho}$) in zebra finch (*Taeniopygia guttata*) across the 18 chromosomes for which we have power to detect recombination hotspots (see *Power to Detect Hotspots*). Both π and $\hat{\rho}$ are shown as rolling means calculated across ten 50 kb windows. To enable comparisons across chromosomes, $\hat{\rho}$ for chromosome Z is shown doubled, because $\rho_{chrZ}=2N_ec$ and $\rho_{autosomes}=4N_ec$ under a simple neutral model and, by an analogous argument, π for chromosome Z is multiplied by $\frac{4}{3}$.

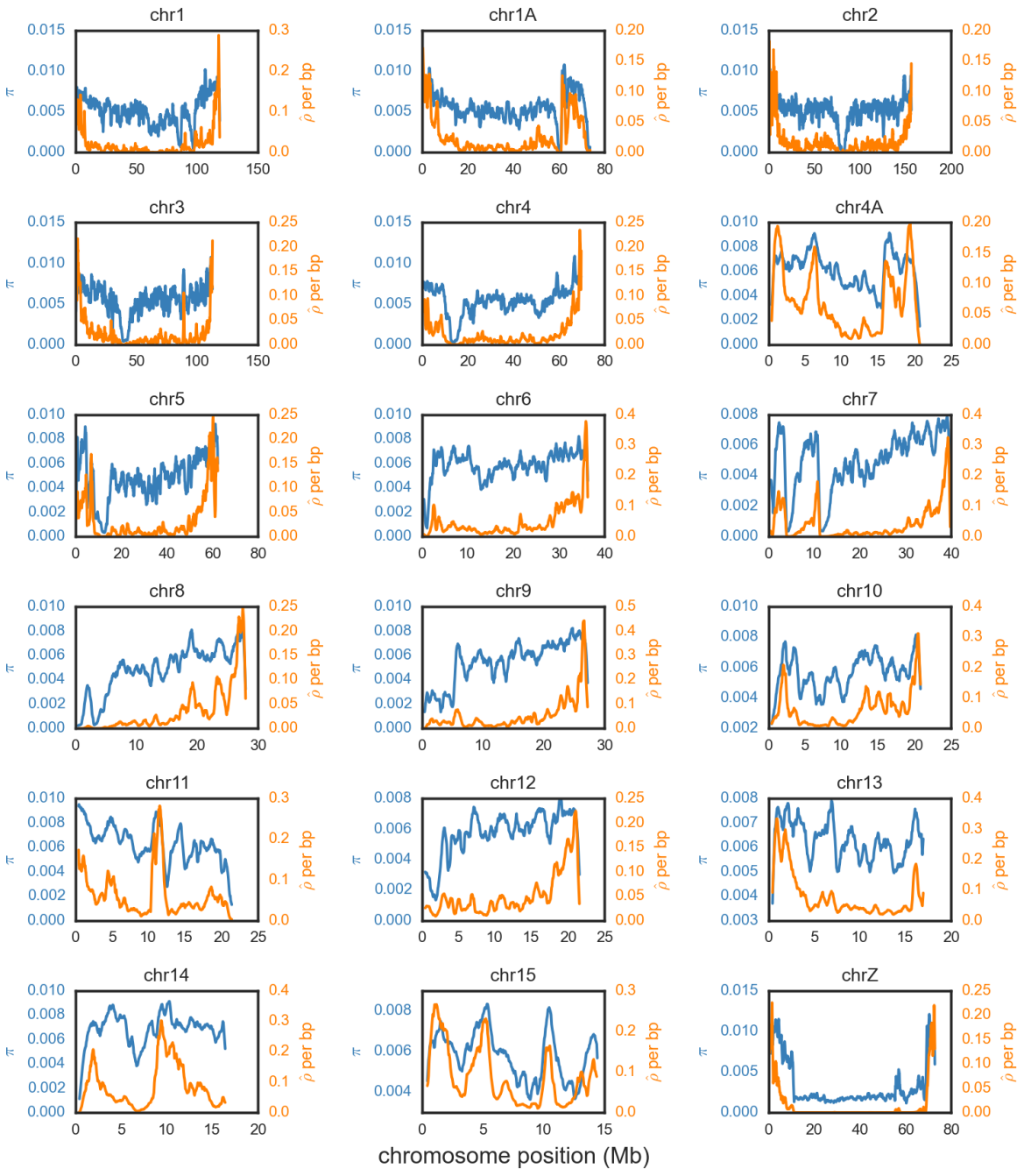

Figure S26: Estimated nucleotide diversity (π) and recombination rate ($\hat{\rho}$) in long-tailed finch (*Poephila acuticauda*) across the 18 chromosomes for which we have power to detect recombination hotspots (see *Power to Detect Hotspots*). Both π and $\hat{\rho}$ are shown as rolling means calculated across ten 50 kb windows. To enable comparisons across chromosomes, $\hat{\rho}$ for chromosome Z is shown doubled, because $\rho_{chrZ}=2N_ec$ and $\rho_{autosomes}=4N_ec$ under a simple neutral model and, by an analogous argument, π for chromosome Z is multiplied by $\frac{4}{3}$.

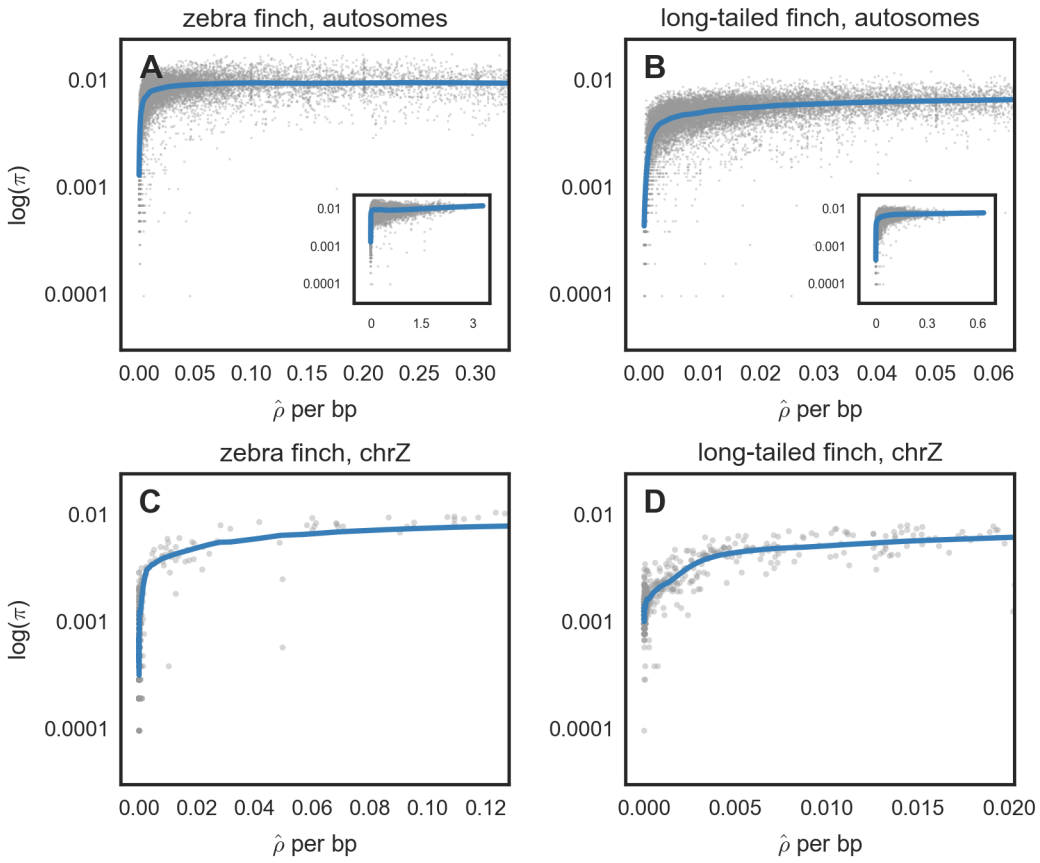

Figure S27: Relationship between nucleotide diversity (π) and recombination rates ($\hat{\rho}$) in zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*); (A-B) The relationship for the 17 largest autosomal chromosomes for which we have power to detect recombination hotspots (see *Comparison to an Existing Genetic Map*) and (C-D) chromosome Z. Rate estimates for chromosome Z should be taken with caution (see *Variant Quality*). Both π and $\hat{\rho}$ were calculated across 50 kb windows with LOESS curves shown for span of 0.1. Insets in the top panels show the full range of $\hat{\rho}$ values.

We note that both π and ρ vary as a function of N_e, so our null expectation is that these two values should be correlated to an extent that depends on the range over which N_e varies across the genome. That said, these results are consistent with a role of linked selection in structuring nucleotide diversity.

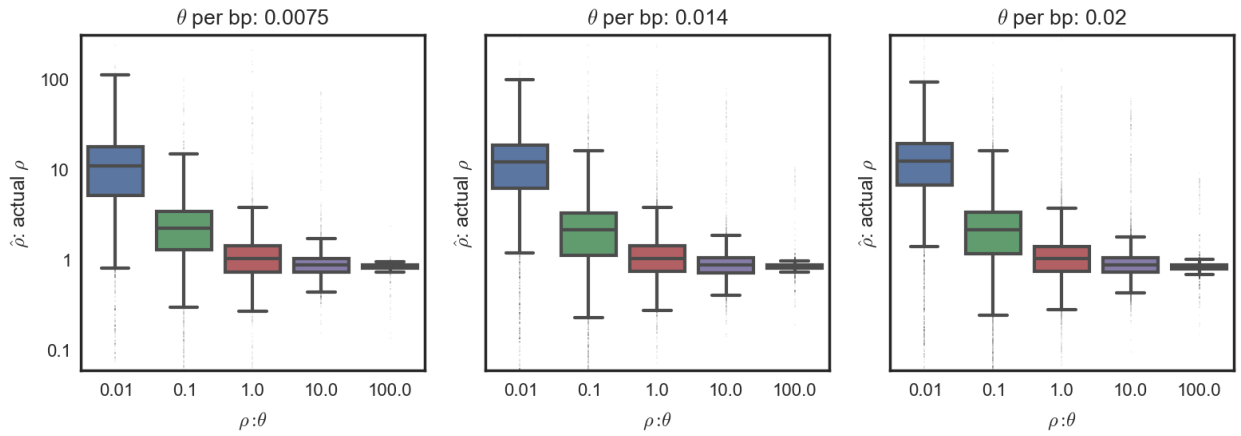

Figure S28: Accuracy of recombination rate (ρ) estimation for different $\rho : \theta$ ratios. Simulations were run for a range of θ that spanned the values seen in zebra finch (*Taeniopygia guttata*) and for ρ that varied four-fold with respect to θ. Recombination rates were estimated using LDhelmet under block penalty 5; and accuracy was calculated as a ratio of $\hat{\rho}$ to ρ.

These simulations help to understand what might happen if exons and introns actually had the same ρ but different θ values, and whether that could lead us to erroneously conclude that ρ is higher in exons because of bias in the estimator. $\hat{\theta}$ (with no singletons) is two times higher in introns than exons in both zebra finch and long-tailed finch, and a lower θ for a given ρ increases the $\rho : \theta$. These simulations show that, for the parameters examined here, increasing the $\rho : \theta$ leads to more accurate estimates of ρ. So, these simulations suggest that, if anything, our ρ estimates in introns should be less accurate than our estimates in exons and that our estimates of ρ in introns would be upwardly biased. This indicates that our result of higher $\hat{\rho}$ in exons versus introns is, if anything, conservative.

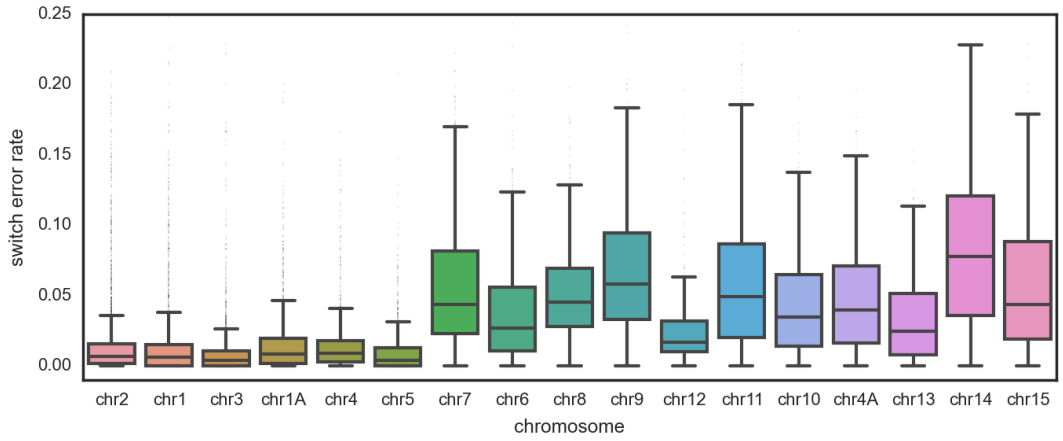

Figure S29: Switch error rates calculated for 50 kb windows for zebra finch (*Taeniopygia guttata*). Chromosomes are shown ordered by their length in descending order. Switch error rates were calculated by comparing haplotypes phased using pedigree information and those phased using phase-informative reads and identifying likely errors in phasing (see *Phasing Errors*).

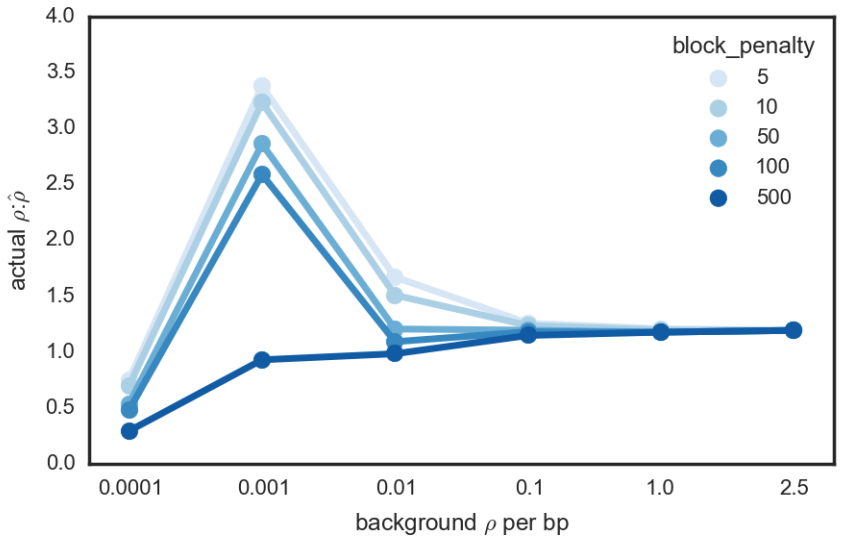

Figure S30: Accuracy of background recombination rate ($\hat{\rho}$) inference under different block penalties. Simulations were run for a range of background ρ that reflects the range seen in zebra finch (*Taeniopygia guttata*); rates were estimated using LDhelmet under a range of block penalties; and the ratio of actual to estimated background ρ was calculated. For each parameter set, 12 Mb of sequence was simulated. Based on these results, we used a block penalty of 100 to generate recombination maps used to characterize broad-scale patterns of recombination.

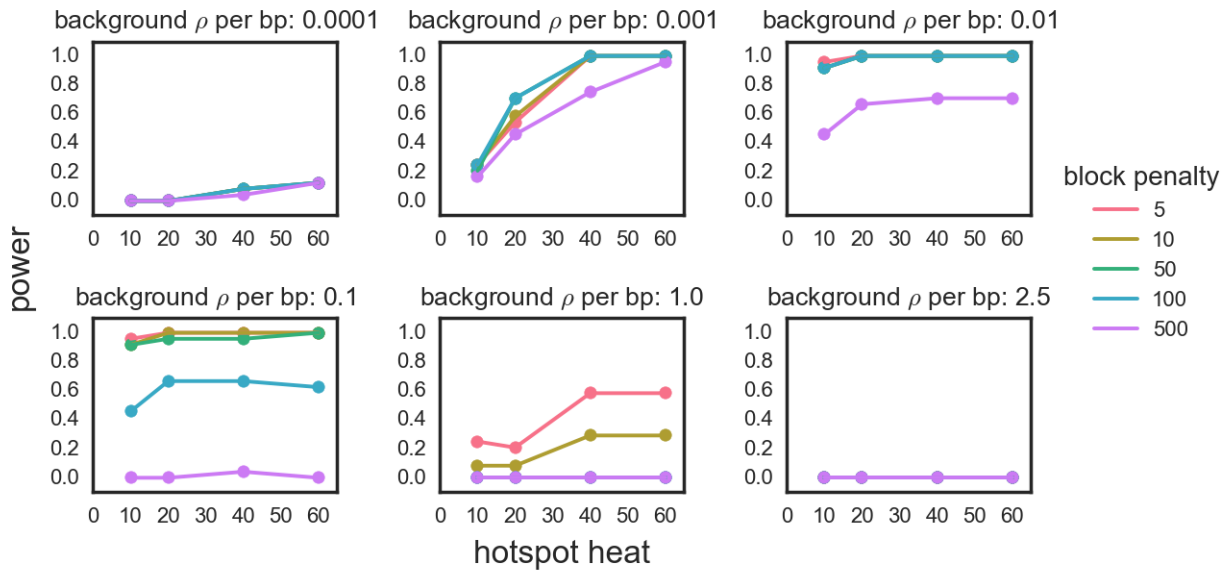

Figure S31: Power to identify hotspots under different block penalties. Simulations were run for a range of background ρ with hotspots of varying heats that reflect the range seen in zebra finch (*Taeniopygia guttata*); rates were estimated using LDhelmet under a range of block penalties; and power was calculated. For each parameter set, 24 hotspots were simulated. When fewer than five lines are visible, lines are overlapping. Based on these results, we used a block penalty of 5 to generate recombination maps used to call hotspots.

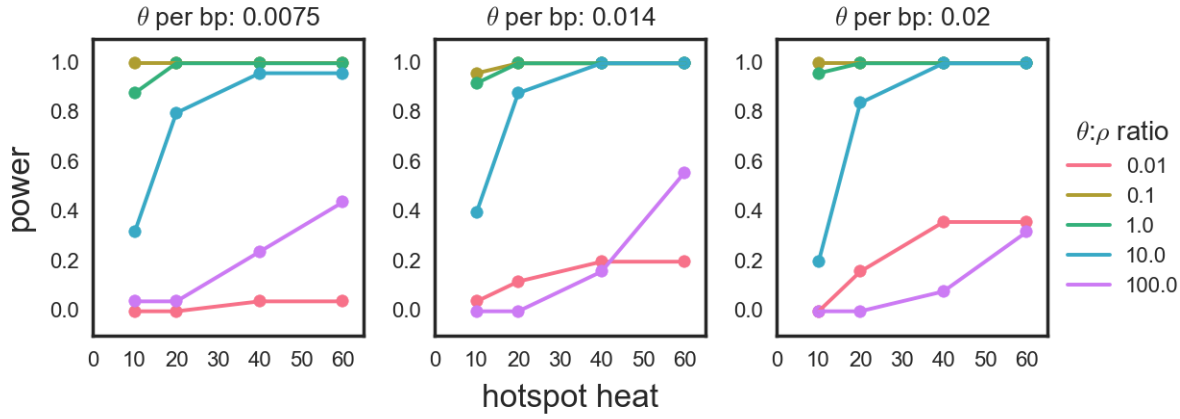

Figure S32: Power to identify hotspots for different $\theta : \rho$ ratios. Simulations were run for a range of θ that spanned the values seen in zebra finch (*Taeniopygia guttata*) and for ρ that varied four-fold with respect to θ. Hotspots were simulated with varying heats; recombination rates were estimated using LDhelmet under block penalty 5; and power to detect hotspots was calculated for each parameter set. For each parameter set, 25 hotspots were simulated.

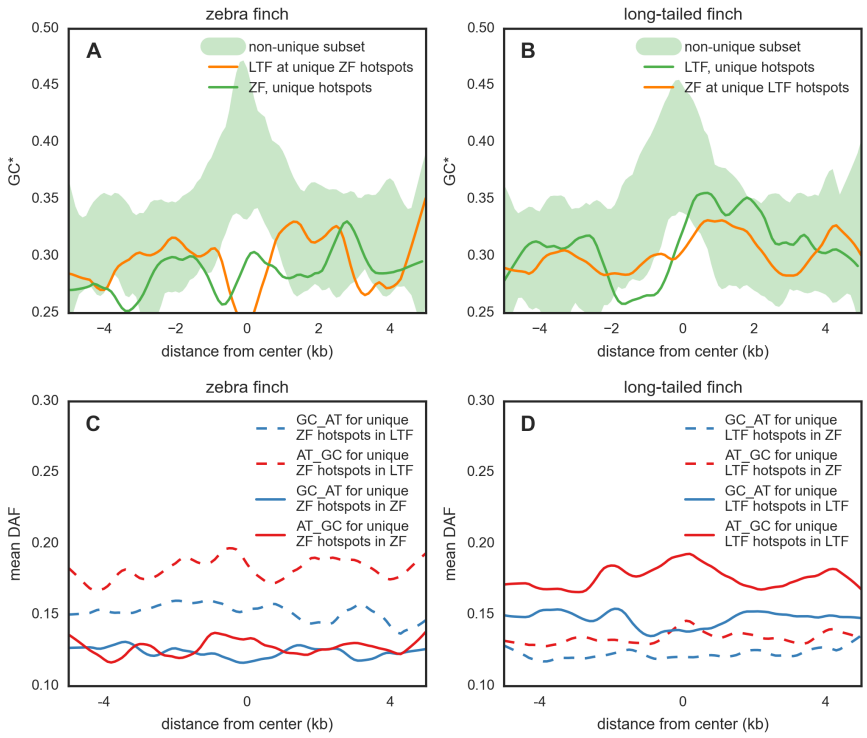

Figure S33: (A - B) Expected equilibrium GC content (GC*) and (C - D) mean derived allele frequency around putatively "unique hotspots", *i.e.*, hotspots that have statistical support in one species and have an estimated heat of 1 in the other species and a likelihood ratio test < 1. 202 such hotspots were inferred in zebra finch (ZF; *Taeniopygia guttata*) and 332 such hotspots were inferred in long-tailed finch (LTF; *Poephila acuticauda)*.

(A - B) Shown is GC* for "unique hotspots" in the species in which they are found and for the same location in the other species in which they are absent; GC* is calculated in 100 bp bins and then smoothed with a LOESS curve for a span of 0.2. Shown in the light blue are the minimum and maximum GC* as calculated from 100 subsamples of the same number of non-unique hotspots. We see no evidence for peaked GC* at these "unique hotspots", and the bootstrap analysis suggests this does not reflect insufficient power.

(C - D) Mean derived allele frequency (DAF) for SNPs of different mutation types around unique hotspots. Mean DAF is calculated in 100 bp bins, sorted by mutation type, and then smoothed with a LOESS curve for a span of 0.2. Variants at all potential CpG sites (where either allele creates a CpG in the ancestral sequence) were excluded because they are more liable to ancestral misidentification. For both species, the mean DAF offers weak support for GC-biased gene conversion at unique hotspots in both the zebra finch and long-tailed finch genome, suggesting that these hotspots are not actually unique.

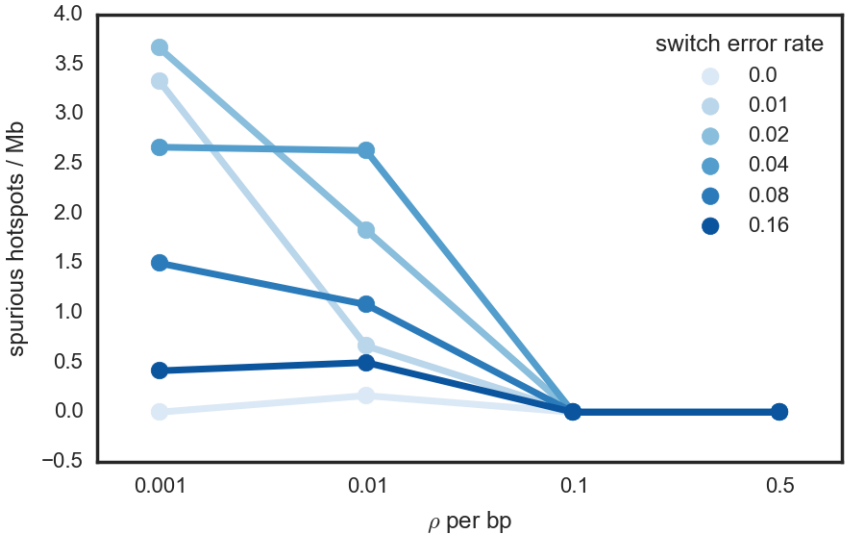

Figure S34: Number of spurious hotspots inferred for different switch error rates. For a range of background ρ, sequences with constant recombination rate (*i.e.*, no hotspots) but varying levels of switch error rates were simulated. Recombination rates were estimated using LDhelmet under block penalty 5, and the number of spurious hotspots found with a heat greater than 5 were counted. For each parameter set, 12 Mb of sequence was simulated. These results suggest that spurious hotspots are most likely at low ρ, across a range of switch error rates.

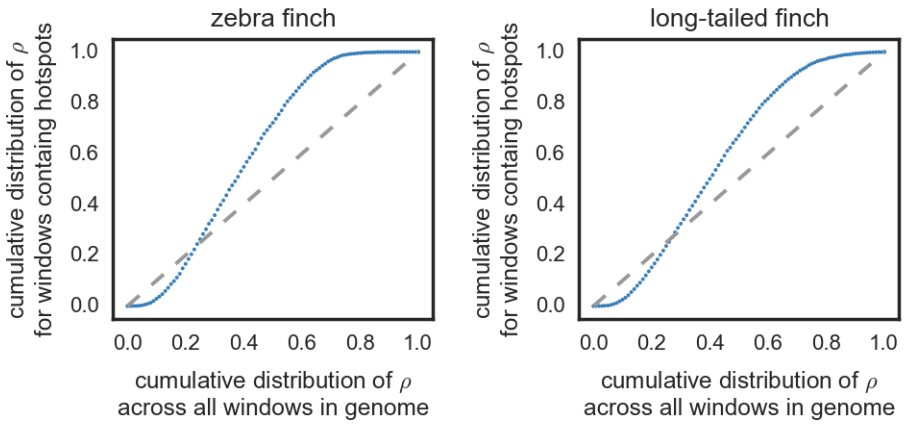

Figure S35: The cumulative distribution of recombination rate variation for 100 kb windows across the whole genome (x-axis) and for 100 kb windows containing hotspots (y-axis) for zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). The actual values are shown as blue points; the line of unity is shown as a dotted gray line. These plots suggest that inferred hotspot locations are not random with respect to background recombination rates. We tend to infer fewer hotspots than expected under random hotspot placement at both low and high recombination rates, which is consistent with power simulations showing reduced power to detect hotspots at both low and high rates (Fig. S4).

3 Supplementary Tables

sample	species	sex	locality	latitude / longitude	Gb of reads mapped	average coverage
G111	*Poephila acuticauda acuticauda*	female	Mount House, Western Australia, Australia	17° 02′ 17.95S, 125° 35′ 49.43E	27.7	22.5
G118	*Poephila acuticauda acuticauda*	male	Mount House, Western Australia, Australia	17° 02′ 17.95S, 125° 35′ 49.43E	31.2	25.3
G163	*Poephila acuticauda acuticauda*	female	Mount House, Western Australia, Australia	17° 02′ 17.95S, 125° 35′ 49.43E	28.1	22.8
G169	*Poephila acuticauda acuticauda*	male	Mount House, Western Australia, Australia	17° 02′ 17.95S, 125° 35′ 49.43E	27.1	22.0
G183	*Poephila acuticauda acuticauda*	male	Mount House, Western Australia, Australia	17° 02′ 17.95S, 125° 35′ 49.43E	30.4	24.6
G250	*Poephila acuticauda acuticauda*	male	Nelson's Hole, Western Australia, Australia	15° 49′ 17.46S, 127° 30′ 23.03E	24.5	19.9
G276	*Poephila acuticauda acuticauda*	male	Nelson's Hole, Western Australia, Australia	15° 49′ 17.46S, 127° 30′ 23.03E	31.4	25.4
G294	*Poephila acuticauda acuticauda*	male	Nelson's Hole, Western Australia, Australia	15° 49′ 17.46S, 127° 30′ 23.03E	24.5	19.9
W2703	*Poephila acuticauda acuticauda*	male	captive-born from parents from Mount House	NA	24.7	20.0
W2994	*Poephila acuticauda acuticauda*	female	captive-born from mother from Mount House and father from Nelson's Hole	NA	28.2	22.8
73783	*Poephila acuticauda hecki*	male	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	23.4	19.0
73788	*Poephila acuticauda hecki*	female	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	22.4	18.1
73790	*Poephila acuticauda hecki*	female	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	31.6	25.6
73900	*Poephila acuticauda hecki*	female	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	24.8	20.1
73903	*Poephila acuticauda hecki*	female	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	34.1	27.6
73907	*Poephila acuticauda hecki*	male	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	26.3	21.3
73933	*Poephila acuticauda hecki*	male	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	27.7	22.4
73942	*Poephila acuticauda hecki*	male	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	28.6	23.2
73948	*Poephila acuticauda hecki*	female	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	32.7	26.5
73958	*Poephila acuticauda hecki*	male	October Creek, Northern Territory, Australia	16° 37′ 52.09S, 134° 51′ 35.69E	38.8	31.4
DBF	*Taeniopygia bichenovii*	male	captive-born	NA	52.4	42.5
26462	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	31° 03′ 55.56S, 141° 50′ 5.88E	34.4	27.9
28339	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	30° 56′ 57.75S, 141° 46′ 2.77E	24.5	19.9
28353	*Taeniopygia guttata*	male	Fowlers Gap, New South Wales, Australia	31° 03′ 55.56S, 141° 50′ 5.88E	40.3	32.7
26721	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	31° 03′ 55.56S, 141° 50′ 5.88E	25.5	20.7
28456	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	31° 03′ 55.56S, 141° 50′ 5.88E	34.2	27.7
28402	*Taeniopygia guttata*	male	Fowlers Gap, New South Wales, Australia	31° 03′ 55.56S, 141° 50′ 5.88E	20.6	16.7
26516	*Taeniopygia guttata*	male	Fowlers Gap, New South Wales, Australia	31° 03′ 55.56S, 141° 50′ 5.88E	27.5	22.3
28404	*Taeniopygia guttata*	male	Fowlers Gap, New South Wales, Australia	31° 01′ 12.13S, 141° 47′ 23.41E	25.8	20.9
26820	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	30° 56′ 57.75S, 141° 46′ 2.77E	36.0	29.2
26733	*Taeniopygia guttata*	male	Fowlers Gap, New South Wales, Australia	31° 01′ 12.13S, 141° 47′ 23.41E	32.2	26.1
28481	*Taeniopygia guttata*	male	Fowlers Gap, New South Wales, Australia	30° 56′ 57.75S, 141° 46′ 2.77E	41.2	33.4
26881	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	31° 01′ 32.58S, 141° 50′ 1.70E	61.9	50.2
26781	*Taeniopygia guttata*	male	Fowlers Gap, New South Wales, Australia	30° 56′ 57.75S, 141° 46′ 2.77E	29.9	24.2
26896	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	30° 56′ 57.75S, 141° 46′ 2.77E	23.0	18.7
26792	*Taeniopygia guttata*	male	Fowlers Gap, New South Wales, Australia	31° 03′ 55.56S, 141° 50′ 5.88E	25.5	20.7
28016	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	31° 03′ 55.56S, 141° 50′ 5.88E	23.1	18.7
26795	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	31° 03′ 55.56S, 141° 50′ 5.88E	42.1	34.2
28078	*Taeniopygia guttata*	female	Fowlers Gap, New South Wales, Australia	30° 56′ 57.75S, 141° 46′ 2.77E	31.4	25.5
28313	*Taeniopygia guttata*	male	Fowlers Gap, New South Wales, Australia	30° 56′ 57.75S, 141° 46′ 2.77E	34.9	28.3
MP1	*Taeniopygia guttata*	female	captive-born; domesticated	NA	33.7	27.3
MP2	*Taeniopygia guttata*	male	captive-born; domesticated	NA	36.7	29.8
MP3	*Taeniopygia guttata*	male	captive-born; domesticated	NA	28.6	23.2
MP4	*Taeniopygia guttata*	male	captive-born; domesticated	NA	24.3	19.7
MP5	*Taeniopygia guttata*	male	captive-born; domesticated	NA	35.7	28.9

Table S1: Information on samples, including their sex, locality, latitude and longitude, Gb of reads mapped, and average coverage. Average coverage was calculated by dividing the length of mapped reads by a genome size of 1.2 Gb (19), which includes the assembled chromosomes and unplaced contigs.

	zebra finch			long-tailed finch			double-barred finch		
	number of sites	SNPs	indels	number of sites	SNPs	indels	number of sites	SNPs	indels
after VQSR	1015266028	50653945	7934954	1015266028	27994027	4373348	1015266028	3216478	458584
after filtering for coverage and repeat-masking	796084847	45600586	6387292	832034824	26177279	3746247	840060375	3029839	359239
after removing Mendelian errors	795865356	45433299	6335088	NA	NA	NA	NA	NA	NA
after masking for putative switch errors	775335201	44629211	6128226	NA	NA	NA	NA	NA	NA
segregating sites	NA	44629211	NA	NA	26177279	NA	NA	3029839	NA
sites used for phasing	NA	42763568	NA	NA	25711643	NA	NA	NA	NA
sites used for estimating recombination rates	NA	20704536	NA	NA	15924774	NA	NA	NA	NA

Table S2: Information on the number of polymorphic SNPs and indels in each population sample of zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). Only biallelic SNPs were used for phasing, and only non-singleton, biallelic SNPs were used to estimate recombination rates. For zebra finch, we additionally excluded 222,428 SNPs from phasing because data were completely missing in the family sample.

	value
number hotspots tested	500
number confirmed for all 3 alternate phasings	469
number confirmed for 2 out of 3 alternate phasings	24
number confirmed for 1 out of 3 alternate phasings	7
mean coefficient of variation of hotspot center across phasings	6.38E-06
mean coefficient of variation of hotspot length across phasings	0.1
mean coefficient of variation of hotspot heat across phasings	0.13

Table S3: Phasing uncertainty and hotspot confirmation. We randomly selected 500 hotspots that had been detected in zebra finch (*Taeniopygia guttata*) and checked if the hotspots were also detected with sequenceLDhot for three alternate, less-likely phasings given by ShapeIt.

	number confirmed
background $\hat{\rho}$, $1\times$	250
background $\hat{\rho}$, $0.5\times$	250
background $\hat{\rho}$, $1.5\times$	211

Table S4: Background recombination rate (ρ) and hotspot confirmation. SequenceLDhot requires the user to set background ρ as known. To determine how mis-specification of ρ affects hotspot inference, we randomly selected 250 hotspots that had been detected in zebra finch (*Taeniopygia guttata*) and determined the number of hotspots confirmed with sequenceLDhot at background ρ $0.5\times$ and $1.5\times$ that of the $\hat{\rho}$ estimated using LDhelmet.

sharing criterion	percent inferred as shared
hotspots overlap over 10% of sequence	68.68%
hotspots overlap over 25% of sequence	68.30%
hotspots overlap over 50% of sequence	64.00%
midpoints within 1 kb	54.50%
midpoints within 2 kb	67.90%
midpoints within 3 kb	72.60%
midpoints within 4 kb	75.60%
midpoints within 5 kb	76.50%

Table S5: Percentage of hotspots inferred as shared between zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*) using different criteria for sharing. Results reported in the main text call hotspots as shared if their midpoints are within 3 kb of each other.

zebra finch				
	A	C	G	T
A	0.455	0.104	0.322	0.119
C	0.206	0.001	0.135	0.659
G	0.659	0.135	0	0.206
T	0.119	0.322	0.103	0.455

long-tailed finch				
	A	C	G	T
A	0.437	0.103	0.344	0.117
C	0.205	0	0.151	0.644
G	0.644	0.151	0	0.205
T	0.117	0.344	0.103	0.436

Table S6: Mutation matrices for zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). These were calculated following (15).

REFERENCES AND NOTES

1. B. de Massy, Initiation of meiotic recombination: How and where? Conservation and specificities among eukaryotes. *Annu. Rev. Genet.* **47**, 563–599 (2013). Medline doi:10.1146/annurev-genet-110711-155423

2. S. Myers, R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, T. S. MacFie, G. McVean, P. Donnelly, Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science* **327**, 876–879 (2010). Medline doi:10.1126/science.1182363

3. I. L. Berg, R. Neumann, K. W. Lam, S. Sarbajna, L. Odenthal-Hesse, C. A. May, A. J. Jeffreys, *PRDM9* variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat. Genet.* **42**, 859–863 (2010). Medline doi:10.1038/ng.658

4. F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, B. de Massy, *PRDM9* is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**, 836–840 (2010). Medline doi:10.1126/science.1183439

5. A. L. Williams, G. Genovese, T. Dyer, N. Altemose, K. Truax, G. Jun, N. Patterson, S. R. Myers, J. E. Curran, R. Duggirala, J. Blangero, D. Reich, M. Przeworski, Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* **4**, e04637 (2015). doi:10.7554/eLife.04637

6. P. L. Oliver, L. Goodstadt, J. J. Bayes, Z. Birtle, K. C. Roach, N. Phadnis, S. A. Beatson, G. Lunter, H. S. Malik, C. P. Ponting, Accelerated evolution of the *Prdm9* speciation gene across diverse metazoan taxa. *PLOS Genet.* **5**, e1000753 (2009). Medline doi:10.1371/journal.pgen.1000753

7. A. G. Hinch, A. Tandon, N. Patterson, Y. Song, N. Rohland, C. D. Palmer, G. K. Chen, K. Wang, S. G. Buxbaum, E. L. Akylbekova, M. C. Aldrich, C. B. Ambrosone, C. Amos, E. V. Bandera, S. I. Berndt, L. Bernstein, W. J. Blot, C. H. Bock, E. Boerwinkle, Q. Cai, N. Caporaso, G. Casey, L. A. Cupples, S. L. Deming, W. R. Diver, J. Divers, M. Fornage, E. M. Gillanders, J. Glessner, C. C. Harris, J. J. Hu, S. A. Ingles, W. Isaacs, E. M. John, W. H. Kao, B. Keating, R. A. Kittles, L. N. Kolonel, E. Larkin, L. Le Marchand, L. H. McNeill, R. C. Millikan, A. Murphy, S. Musani, C. Neslund-Dudas, S. Nyante, G. J. Papanicolaou, M. F. Press, B. M. Psaty, A. P. Reiner, S. S. Rich, J. L. Rodriguez-Gil, J. I. Rotter, B. A. Rybicki, A. G. Schwartz, L. B. Signorello, M. Spitz, S. S. Strom, M. J. Thun, M. A. Tucker, Z. Wang, J. K. Wiencke, J. S. Witte, M. Wrensch, X. Wu, Y. Yamamura, K. A. Zanetti, W. Zheng, R. G. Ziegler, X. Zhu, S. Redline, J. N. Hirschhorn, B. E. Henderson, H. A. Taylor Jr., A. L. Price, H. Hakonarson, S. J. Chanock, C. A. Haiman, J. G. Wilson, D. Reich, S. R. Myers, The landscape of recombination in African Americans. *Nature* **476**, 170–175 (2011). Medline doi:10.1038/nature10336

8. L. S. Stevison, A. E. Woerner, J. M. Kidd, J. L. Kelley, K. R. Veeramah, K. F. McManus, C. D. Bustamante, M. F. Hammer, J. D. Wall, The time-scale of recombination rate evolution in great apes. http://biorxiv.org/search/013755 (2015).

9. A. Auton, A. Fledel-Alon, S. Pfeifer, O. Venn, L. Ségurel, T. Street, E. M. Leffler, R. Bowden, I. Aneas, J. Broxholme, P. Humburg, Z. Iqbal, G. Lunter, J. Maller, R. D.

Hernandez, C. Melton, A. Venkat, M. A. Nobrega, R. Bontrop, S. Myers, P. Donnelly, M. Przeworski, G. McVean, A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**, 193–198 (2012). Medline doi:10.1126/science.1216872

10. K. Brick, F. Smagulova, P. Khil, R. D. Camerini-Otero, G. V. Petukhova, Genetic recombination is directed away from functional genomic elements in mice. *Nature* **485**, 642–645 (2012). Medline doi:10.1038/nature11089

11. A. Auton, Y. R. Li, J. Kidd, K. Oliveira, J. Nadel, J. K. Holloway, J. J. Hayward, P. E. Cohen, J. M. Greally, J. Wang, C. D. Bustamante, A. R. Boyko, Genetic recombination is targeted towards gene promoter regions in dogs. *PLOS Genet.* **9**, e1003984 (2013).

12. E. Axelsson, M. T. Webster, A. Ratnakumar, C. P. Ponting, K. Lindblad-Toh; LUPA Consortium, Death of *PRDM9* coincides with stabilization of the recombination landscape in the dog genome. *Genome Res.* **22**, 51–63 (2012). Medline doi:10.1101/gr.124123.111

13. J. Pan, M. Sasaki, R. Kniewel, H. Murakami, H. G. Blitzblau, S. E. Tischfield, X. Zhu, M. J. Neale, M. Jasin, N. D. Socci, A. Hochwagen, S. Keeney, A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* **144**, 719–731 (2011). Medline doi:10.1016/j.cell.2011.02.009

14. K. Choi, X. Zhao, K. A. Kelly, O. Venn, J. D. Higgins, N. E. Yelina, T. J. Hardcastle, P. A. Ziolkowski, G. P. Copenhaver, F. C. Franklin, G. McVean, I. R. Henderson, *Arabidopsis* meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat. Genet.* **45**, 1327–1336 (2013). Medline doi:10.1038/ng.2766

15. A. H. Chan, P. A. Jenkins, Y. S. Song, Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLOS Genet.* **8**, e1003090 (2012). Medline

16. A. Wallberg, S. Glémin, M. T. Webster. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLOS Genet.* **11**, e1005189 (2015).

17. T. Kaur, M. V. Rockman, Crossover heterogeneity in the absence of hotspots in *Caenorhabditis elegans*. *Genetics* **196**, 137–148 (2014). Medline doi:10.1534/genetics.113.158857

18. Materials and methods are available as supplementary materials on *Science* Online.

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

20. E. M. Leffler, K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel, A. Venkat, P. Andolfatto, M. Przeworski, Revisiting an old riddle: What determines genetic diversity levels within species? *PLOS Biol.* **10**, e1001388 (2012). Medline doi:10.1371/journal.pbio.1001388

21. N. Backström, W. Forstmeier, H. Schielzeth, H. Mellenius, K. Nam, E. Bolund, M. T. Webster, T. Ost, M. Schneider, B. Kempenaers, H. Ellegren, The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Res.* **20**, 485–495 (2010). Medline doi:10.1101/gr.101410.109

22. C. C. Weber, B. Boussau, J. Romiguier, E. D. Jarvis, H. Ellegren, Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol.* **15**, 549 (2014). Medline doi:10.1186/s13059-014-0549-1

23. G. Zhang, P. Parker, B. Li, H. Li, J. Wang, The genome of Darwin's Finch (*Geospiza fortis*). *GigaScience* 10.5524/100040 (2012).

24. T. Kawakami, L. Smeds, N. Backström, A. Husby, A. Qvarnström, C. F. Mugal, P. Olason, H. Ellegren, A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Mol. Ecol.* **23**, 4035–4058 (2014). Medline doi:10.1111/mec.12810

25. I. Lam, S. Keeney, Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science* **350**, 923–937 (2015).

26. A. Nicolas, D. Treco, N. P. Schultes, J. W. Szostak, An initiation site for meiotic gene conversion in the yeast *Saccharomyces cerevisiae*. *Nature* **338**, 35–39 (1989). Medline doi:10.1038/338035a0

27. U. Hellsten, K. M. Wright, J. Jenkins, S. Shu, Y. Yuan, S. R. Wessler, J. Schmutz, J. H. Willis, D. S. Rokhsar, Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19478–19482 (2013). Medline doi:10.1073/pnas.1319032110

28. P. A. Jones, Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012). Medline doi:10.1038/nrg3230

29. N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, E. Segal, The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009). Medline doi:10.1038/nature07667

30. A. M. Deaton, A. Bird, CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011). Medline doi:10.1101/gad.2037511

31. E. Wijnker, G. Velikkakam James, J. Ding, F. Becker, J. R. Klasen, V. Rawat, B. A. Rowan, D. F. de Jong, C. B. de Snoo, L. Zapata, B. Huettel, H. de Jong, S. Ossowski, D. Weigel, M. Koornneef, J. J. Keurentjes, K. Schneeberger, The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *eLife* **2**, e01426 (2013). Medline doi:10.7554/eLife.01426

32. L. Christidis, Chromosomal evolution within the family Estrildidae (Aves) I. The Poephilae. *Genetica* **71**, 81–97 (1986). doi:10.1007/BF00058691

33. Y. Itoh, K. Kampf, C. N. Balakrishnan, A. P. Arnold, Karyotypic polymorphism of the zebra finch Z chromosome. *Chromosoma* **120**, 255–264 (2011). Medline doi:10.1007/s00412-010-0308-3

34. B. L. Dumont, B. A. Payseur, Evolution of the genomic rate of recombination in mammals. *Evolution* **62**, 276–294 (2008). Medline doi:10.1111/j.1558-5646.2007.00278.x

35. M. I. Jensen-Seaman, T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C. F. Chen, M. A. Thomas, D. Haussler, H. J. Jacob, Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14**, 528–538 (2004). Medline doi:10.1101/gr.1970304

36. S. P. Otto, J. R. Pannell, C. L. Peichel, T. L. Ashman, D. Charlesworth, A. K. Chippindale, L. F. Delph, R. F. Guerrero, S. V. Scarpino, B. F. McAllister, About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet.* **27**, 358–367 (2011). Medline doi:10.1016/j.tig.2011.05.001

37. M. W. Nachman, Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **17**, 481–485 (2001). Medline doi:10.1016/S0168-9525(01)02409-X

38. D. J. Begun, C. F. Aquadro, Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992). Medline doi:10.1038/356519a0

39. C. F. Mugal, P. F. Arndt, H. Ellegren, Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. *Mol. Biol. Evol.* **30**, 1700–1712 (2013). Medline doi:10.1093/molbev/mst067

40. B. Charlesworth, M. T. Morgan, D. Charlesworth, The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993). Medline

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

42. S. C. Griffith, S. R. Pryke, M. Mariette, Use of nest-boxes by the Zebra Finch (*Taeniopygia guttata*): Implications for reproductive success and research. *Emu* **108**, 311 (2009). doi:10.1071/MU08033

43. L. A. Rollins, N. Svedin, S. R. Pryke, S. C. Griffith, The role of the Ord Arid Intrusion in the historical and contemporary genetic division of long-tailed finch subspecies in northern Australia. *Ecol. Evol.* **2**, 1208 (2012).

44. G. Pesole, F. Mignone, C. Gissi, G. Grillo, F. Licciulli, S. Liuni, Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* **276**, 73–81 (2001). Medline doi:10.1016/S0378-1119(01)00674-6

45. J. Nylander, MrAIC (2004); https://github.com/nylander/MrAIC.

46. S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010). Medline doi:10.1093/sysbio/syq010

47. L. Liu, L. Yu, D. K. Pearl, S. V. Edwards, Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* **58**, 468–477 (2009). Medline doi:10.1093/sysbio/syp031

48. R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C. H. Wu, D. Xie, M. A. Suchard, A. Rambaut, A. J. Drummond, BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014). Medline doi:10.1371/journal.pcbi.1003537

49. R. J. Agate, B. B. Scott, B. Haripal, C. Lois, F. Nottebohm, Transgenic songbirds offer an opportunity to develop a genetic model for vocal learning. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 17963–17967 (2009). Medline doi:10.1073/pnas.0909139106

50. D. M. Hooper, T. D. Price, Rates of karyotypic evolution in Estrildid finches differ between island and continental clades. *Evolution* **69**, 890–903 (2015). Medline doi:10.1111/evo.12633

51. T. D. Price, D. M. Hooper, C. D. Buchanan, U. S. Johansson, D. T. Tietze, P. Alström, U. Olsson, M. Ghosh-Harihar, F. Ishtiaq, S. K. Gupta, J. Martens, B. Harr, P. Singh, D. Mohan, Niche filling slows the diversification of Himalayan songbirds. *Nature* **509**, 222–225 (2014). Medline doi:10.1038/nature13272

52. W. Jetz, G. H. Thomas, J. B. Joy, K. Hartmann, A. O. Mooers, The global diversity of birds in space and time. *Nature* **491**, 444–448 (2012). Medline doi:10.1038/nature11631

53. E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F. Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, M. S. Bayzid, V. Zavidovych, S. Subramanian, T. Gabaldón, S. Capella-Gutiérrez, J. Huerta-Cepas, B. Rekepalli, K. Munch, M. Schierup, B. Lindow, W. C. Warren, D. Ray, R. E. Green, M. W. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Y. Huang, E. P. Derryberry, M. F. Bertelsen, F. H. Sheldon, R. T. Brumfield, C. V. Mello, P. V. Lovell, M. Wirthlin, M. P. Schneider, F. Prosdocimi, J. A. Samaniego, A. M. Vargas Velazquez, A. Alfaro-Núñez, P. F. Campos, B. Petersen, T. Sicheritz-Ponten, A. Pas, T. Bailey, P. Scofield, M. Bunce, D. M. Lambert, Q. Zhou,

P. Perelman, A. C. Driskell, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang, J. Wang, L. Smeds, F. E. Rheindt, M. Braun, J. Fjeldsa, L. Orlando, F. K. Barker, K. A. Jønsson, W. Johnson, K. P. Koepfli, S. O'Brien, D. Haussler, O. A. Ryder, C. Rahbek, E. Willerslev, G. R. Graves, T. C. Glenn, J. McCormack, D. Burt, H. Ellegren, P. Alström, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft, E. L. Braun, T. Warnow, W. Jun, M. T. Gilbert, G. Zhang, Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014). Medline

54. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genet.* **8**, e1002967 (2012). Medline doi:10.1371/journal.pgen.1002967

55. D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian population history. *Nature* **461**, 489–494 (2009). Medline doi:10.1038/nature08365

56. J. A. St John, E. L. Braun, S. R. Isberg, L. G. Miles, A. Y. Chong, J. Gongora, P. Dalzell, C. Moran, B. Bed'hom, A. Abzhanov, S. C. Burgess, A. M. Cooksey, T. A. Castoe, N. G. Crawford, L. D. Densmore, J. C. Drew, S. V. Edwards, B. C. Faircloth, M. K. Fujita, M. J. Greenwold, F. G. Hoffmann, J. M. Howard, T. Iguchi, D. E. Janes, S. Y. Khan, S. Kohno, A. J. de Koning, S. L. Lance, F. M. McCarthy, J. E. McCormack, M. E. Merchant, D. G. Peterson, D. D. Pollock, N. Pourmand, B. J. Raney, K. A. Roessler, J. R. Sanford, R. H. Sawyer, C. J. Schmidt, E. W. Triplett, T. D. Tuberville, M. Venegas-Anaya, J. T. Howard, E. D. Jarvis, L. J. Guillette Jr., T. C. Glenn, R. E. Green, D. A. Ray, Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biol.* **13**, 415 (2012). Medline

57. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). Medline doi:10.1186/1471-2105-10-421

58. E. Birney, M. Clamp, R. Durbin, GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004). Medline doi:10.1101/gr.1865504

59. F. Baudat, Y. Imai, B. de Massy, Meiotic recombination in mammals: Localization and regulation. *Nat. Rev. Genet.* **14**, 794–806 (2013). Medline doi:10.1038/nrg3573

60. W. J. Kent, BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002). Medline doi:10.1101/gr.229202

61. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). Medline doi:10.1093/bioinformatics/btq033

62. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). Medline doi:10.1101/gr.107524.110

63. Q. Zhou, J. Zhang, D. Bachtrog, N. An, Q. Huang, E. D. Jarvis, M. T. Gilbert, G. Zhang, Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346**, 1246338 (2014). Medline doi:10.1126/science.1246338

64. L. Smeds, T. Kawakami, R. Burri, P. Bolivar, A. Husby, A. Qvarnström, S. Uebbing, H. Ellegrena, Genomic identification and characterization of the pseudoautosomal region in highly differentiated avian sex chromosomes. *Nat. Commun.* **5**, 5448 (2014).

65. R. S. Harris, thesis, The Pennsylvania State University, State College, PA (2007).

66. G. Lunter, M. Goodson, Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011). Medline doi:10.1101/gr.111120.110

67. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. http://arxiv.org/abs/1303.3997 (2013).

68. Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, G. McVean, De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012). Medline doi:10.1038/ng.1028

69. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). Medline doi:10.1093/bioinformatics/btp352

70. 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010). Medline

71. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). Medline doi:10.1086/519795

72. J. K. Pritchard, M. Przeworski, Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001). Medline doi:10.1086/321275

73. C. M. Rands, A. Darling, M. Fujita, L. Kong, M. T. Webster, C. Clabaut, R. D. Emes, A. Heger, S. Meader, M. B. Hawkins, M. B. Eisen, C. Teiling, J. Affourtit, B. Boese, P. R. Grant, B. R. Grant, J. A. Eisen, A. Abzhanov, C. P. Ponting, Insights into the evolution of Darwin's finches from comparative analysis of the *Geospiza magnirostris* genome sequence. *BMC Genomics* **14**, 95 (2013). Medline doi:10.1186/1471-2164-14-95

74. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). Medline doi:10.1038/nmeth.1923

75. G. A. Watterson, On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975). Medline doi:10.1016/0040-5809(75)90020-9

76. F. Tajima, The effect of change in population size on DNA polymorphism. *Genetics* **123**, 597–601 (1989). Medline

77. C. N. Balakrishnan, S. V. Edwards, Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (*Taeniopygia guttata*). *Genetics* **181**, 645–660 (2009). Medline doi:10.1534/genetics.108.094250

78. A. L. Williams, D. E. Housman, M. C. Rinard, D. K. Gifford, Rapid haplotype inference for nuclear families. *Genome Biol.* **11**, R108 (2010). Medline doi:10.1186/gb-2010-11-10-r108

79. O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, J. Marchini, Haplotype estimation using sequencing reads. *AJHG* **93**, 687–696 (2013). doi:10.1016/j.ajhg.2013.09.002

80. M. Stephens, P. Donnelly, A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162–1169 (2003). Medline doi:10.1086/379378

81. G. Coop, M. Przeworski, An evolutionary view of human recombination. *Nat. Rev. Genet.* **8**, 23–34 (2007). Medline doi:10.1038/nrg1947

82. K. Choi, I. R. Henderson, Meiotic recombination hotspots - a comparative view. *Plant J.* **83**, 52–61 (2015).

83. A. Auton, G. McVean, Recombination rate estimation in the presence of hotspots. *Genome Res.* **17**, 1219–1227 (2007). Medline doi:10.1101/gr.6386707

84. G. K. Chen, P. Marjoram, J. D. Wall, Fast and flexible simulation of DNA sequence data. *Genome Res.* **19**, 136–142 (2009). Medline doi:10.1101/gr.083634.108

85. P. Fearnhead, SequenceLDhot: Detecting recombination hotspots. *Bioinformatics* **22**, 3061–3066 (2006). Medline doi:10.1093/bioinformatics/btl540

86. The International HapMap Consortium, A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005). Medline doi:10.1038/nature04226

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

88. N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. U.S.A.* **48**, 582–592 (1962). Medline doi:10.1073/pnas.48.4.582

89. J. Meunier, L. Duret, Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21**, 984–990 (2004). Medline doi:10.1093/molbev/msh070

90. T. L. Bailey, J. Johnson, C. E. Grant, W. S. Noble, The MEME Suite. *Nucl. Acids Res.* **43**, W39–W49 (2015).

91. P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. Singh Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suárez, J. Harrow, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, S. M. J. Searle, Ensembl 2012. *Nucl. Acids Res.* **40**, D84–D90 (2011).

92. A. Roberts, L. Pachter, Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013). Medline doi:10.1038/nmeth.2251

93. T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, J. O. Korbel, DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012). Medline doi:10.1093/bioinformatics/bts378

94. F. Pratto, K. Brick, P. Khil, F. Smagulova, G. V. Petukhova, R. D. Camerini-Otero, Recombination initiation maps of individual human genomes. *Science* **346**, 1256442 (2014). Medline doi:10.1126/science.1256442