PNAS RESEARCH ARTICLE | EVOLUTION

Fitness consequences of structural variation inferred from a House Finch pangenome

Bohao Fang[a,b,1] and Scott V. Edwards[a,b,1]

Edited by Catherine Peichel, Universitat Bern, Bern, Switzerland; received May 17, 2024; accepted October 3, 2024

Genomic structural variants (SVs) play a crucial role in adaptive evolution, yet their average fitness effects and characterization with pangenome tools are understudied in wild animal populations. We constructed a pangenome for House Finches (*Haemorhous mexicanus*), a model for studies of host-pathogen coevolution, using long-read sequence data on 16 individuals (32 de novo-assembled haplotypes) and one outgroup. We identified 887,118 SVs larger than 50 base pairs, mostly (60%) involving repetitive elements, with reduced SV diversity in the eastern US as a result of its introduction by humans. The distribution of fitness effects of genome-wide SVs was estimated using maximum likelihood approaches and revealed that SVs in both coding and noncoding regions were on average more deleterious than smaller indels or single nucleotide polymorphisms. The reference-free pangenome facilitated identification of a > 10-My-old, 11-megabase-long pericentric inversion on chromosome 1. We found that the genotype frequencies of the inversion, estimated from 135 birds widely sampled temporally and geographically, increased steadily over the 25 y since House Finches were first exposed to the bacterial pathogen *Mycoplasma gallisepticum* and showed signatures of balancing selection, capturing genes related to immunity and telomerase activity. We also observed shorter telomeres in populations with a greater number of years exposure to *Mycoplasma*. Our study illustrates the utility of long-read sequencing and pangenome methods for understanding wild animal populations, estimating fitness effects of genome-wide SVs, and advancing our understanding of adaptive evolution through structural variation.

transposable element | gene regulation | inversion | bird | population genomics

Significance

Prevailing genomic research on adaptive and neutral evolution has focused primarily on single nucleotide polymorphisms. However, structural variation (SV) plays a critical role in animal adaptive evolution, often directly underlying fitness-relevant traits, although their average effects on fitness are less well understood. Our study constructs a pangenome for the House Finch using long-read sequencing, capturing the full spectrum of genomic diversity without use of a reference genome. In addition to detecting over 800,000 SVs, we also document a large inversion that shows evidence of contributing to disease resistance. Our use of long-read sequencing and pangenomic approaches in a wild bird population presents a compelling approach to understanding the complexities of molecular ecology and adaptive evolution.

Structural variants (SVs) are large DNA changes in the genome, consisting of insertions, deletions, and rearrangements such as inversions and translocations. They differ from the more common and simpler "single nucleotide polymorphisms" (SNPs), which involve only single DNA base changes. Some SVs have been shown to be important for adaptation in natural populations (1–6), influencing morphology (7–9), behavior (10–12), and other physiological traits, such as disease resistance in humans (13), animals (14, 15), and plants (16). However, compared to SNPs, we know little about the average fitness effects of SVs, due in part to their underrepresentation in short-read sequencing data and understudied links to phenotypes (3, 17–19).

One useful way of documenting SVs is via pangenomes. A pangenome models the complete set of genomic elements found within a species or clade, in contrast to reference-based methods, which compare sequences to a single genome, potentially biasing studies and missing variation due to choice of reference (20). Using long-read DNA sequencing methods, the totality of genetic variation in genomes can now be better summarized in an organism's pangenome, encompassing not only SNPs but also simple and complex SVs (21, 22). Pangenome graphs, including de Bruijn and variation graphs, are considered superior data structures for categorizing extensive haplotypic genetic diversity within a single, data-rich model (20, 23–27). For instance, pangenome graphs have demonstrably improved mapping and variant discovery rates compared to reference genomes, allowing more comprehensive analysis of SVs (21, 28, 29). The application of pangenomic approaches is currently confined to humans (21, 29, 30), agricultural crops (31) livestock (32–36), and microbes (37, 38). Pangenomics in wild animals remains limited (39, 40), thus far mostly restricted by small sample sizes or focused on subspecies or higher taxonomic levels, making it difficult to evaluate the fitness effects of SVs.

The House Finch (*Haemorhous mexicanus*) is a model organism for studying host-pathogen coevolution; in the mid-1990s it experienced an epizootic involving a conjunctivitis-causing bacterial pathogen called *Mycoplasma gallisepticum* (MG) (41–46). House finches are native to the western US and were introduced to the eastern US by humans around 1940, where they adapted rapidly and began to expand west toward their original range (47). Following an MG outbreak in the eastern US from 1994 to 1998 and

Author affiliations: [a]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138; and [b]Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138

Author contributions: B.F. and S.V.E. designed research; B.F. performed research; B.F. and S.V.E. contributed new reagents/analytic tools; B.F. analyzed data; and B.F. and S.V.E. wrote the paper.

Competing interest statement: S.V.E. is an editor for PNAS and is not involved in any of the editorial decisions regarding this manuscript.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: sedwards@fas.harvard.edu or bfang@fas.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2409943121/-/DCSupplemental.

Published November 12, 2024.

a subsequent outbreak in the western US in 2002, their resistance to the disease appears to have increased (48) as measured by gene expression (42, 49, 50), antibodies (51, 52), survival experiments (53, 54), and population surveys (48, 55). Here, we present a pangenome of the House Finch based on high-quality haplotypes assembled from 17 birds to characterize the full spectrum of genomic variants, including SVs, and investigate the fitness effects of SVs and their associations, if any, with adaptive resistance to MG.

Results

House Finch Pangenome Captures Genome-Wide Structural Variation. To construct the pangenome, we generated de novo assemblies for 16 House Finch samples and an outgroup, the Common Rosefinch [*Carpodacus erythrinus*; diverged 12.9 Mya (56, 57)], using PacBio HiFi long-read sequencing at approximately 42× coverage per bird (Datasets S1 and S2). We selected eight samples each from the western and eastern US (western and eastern, respectively) for a balanced genetic representation of the two populations (58) (Fig. 1*B*; *Materials and Methods*). The rationale for selecting the Common Rosefinch as the outgroup is detailed in *SI Appendix*, *Methods*. We also incorporated a chromosome-level assembly from a House Finch collected in California, produced and curated by the Vertebrate Genomes Project (VGP genome: 39 autosomes and sex chromosomes Z and W), primarily to provide a set of stable genomic coordinates in the pangenome and downstream analyses by us and other researchers. Our annotation strategies, which included in silico and evidence-based approaches (*Materials and Methods*), identified 22,080 protein-coding genes and 18.1% repeat content (*SI Appendix*, Fig. S1) across the VGP genome. Two haplotypes per sample were assembled with hifiasm (59) (Fig. 1*C*; quality metrics in Dataset S2). All data have been made publicly available (Data availability).

The 35 haplotype assemblies, composed of 32 House Finch haplotypes, two Common Rosefinch haplotypes, and the VGP genome assembly, were used for pangenome construction. We built a separate pangenome graph for each of the 39 autosomes using the PanGenome Graph Builder (PGGB) (25), a pipeline for constructing unbiased pangenome graphs using all-to-all genome alignments without relying on a reference genome. The selection of PGGB as the primary pangenome graph builder was based on its ability to represent comprehensively genomic variants of all sizes and its prior application in large-scale genomic projects (21) (detailed justification in *Materials and Methods*). In the graph model, nodes represent DNA segments. Each node can orient in two ways: forward or reverse, creating a bidirected graph with four potential edges between any node pair to represent all combinations of orientations (*SI Appendix* Fig. S2). Haplotype sequences are depicted as paths within this graph. To characterize variants in the graph, we decomposed the graph to identify "bubble" subgraphs corresponding to nonoverlapping variants (*SI Appendix*,

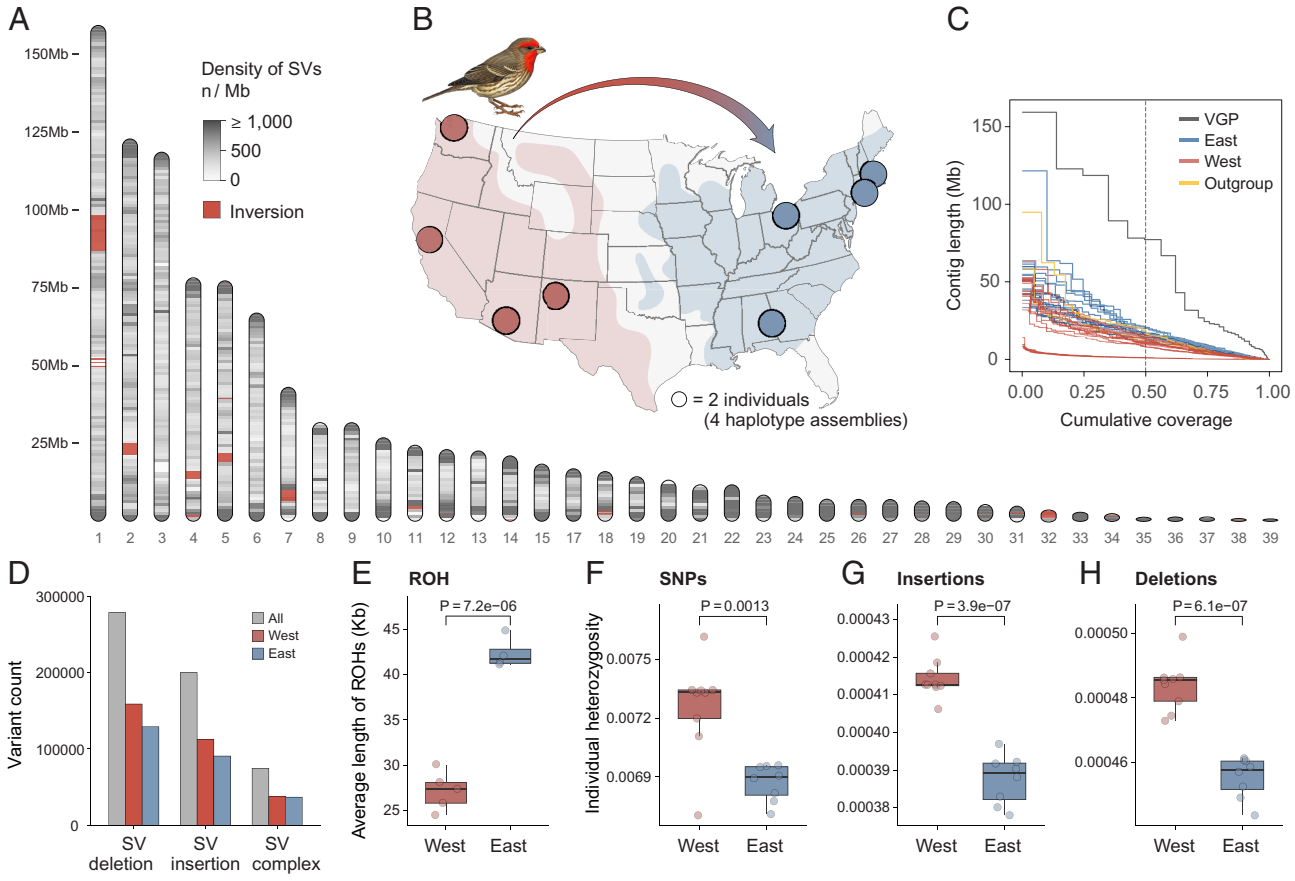

Fig. 1. Genome-wide structural variants captured in the House Finch pangenome. (*A*) Genome-wide density of SVs larger than 50 bp. Sex chromosomes and chromosome 16 are not included in the analyses (*Materials and Methods*). (*B*) Geographic distribution of 16 sampled House Finches. (*C*) Assembly contiguity of the 32 haplotype assemblies, a chromosome-level assembly from the VGP, and two haplotype assemblies of the outgroup species Common Rosefinch (*C. erythrinus*), shown in a NGx graph representing the contig length at which x% of the genome is covered. (*D*) Quantity of SVs in House Finch populations extracted from the PGGB pangenome graph. (*E*) Population-wise runs of homozygosity (ROH). (*F*–*H*) Individual heterozygosity for SNPs (*F*), insertions (*G*), and deletions (*H*). Statistical significance (*P*-values) was determined using two-tailed *t* tests. Bird illustration by Norman Arlott, © Lynx Edicions.

Fig. S2). In subsequent summaries of graph and variant statistics, we excluded the sex chromosomes and chromosome 16, which is composed of 93% repeats (justification in *SI Appendix*, *Methods*) and was fragmented even in the VGP assembly. Across 38 autosomes other than chromosome 16, the House Finch pangenome graph (including the outgroup) consists of 203,515,248 nodes, 286,630,623 edges, and 29,517 paths (see Dataset S3 for graph statistics per chromosome).

Decomposing the PGGB pangenome graph, we classified variants into 26,470,883 SNPs across House Finch haplotypes (*Materials and Methods*); 4,535,856 INDELs (<50 base pairs [bp]) including 1,411,408 biallelic insertions (INS) polarized relative to the outgroup, 2,184,433 deletions (DEL), 368,780 complex INDELs (INDEL-complex); and 887,118 SVs (≥50 bp; Fig. 1*D*), including 199,924 SV-insertions (SV-INS), 278,350 SV-deletions (SV-DEL), and 74,609 complex SVs (SV-complex). Complex variants are defined as sites with different allelic sizes that are not 1 bp in length (*Materials and Methods*). Multiallelic variants (361,830 SNPs, 571,235 INDELs, and 334,235 SVs) were also identified (Dataset S4). To enhance discovery of inversions and other SVs, and to compare the results of different methods for detecting SVs, we additionally used minigraph (23), a pangenome graph builder tailored to identify large SVs, as well as SVIM-asm (60) and SyRi (61), the latter two being reference-based SV callers (*Materials and Methods*; Dataset S5). Using these tools we identified 343 inversions ranging in size from 50 bp to 11.3 megabases (Mb) (Fig. 1*A*; Dataset S5); 163 inversions (48%) were identified by at least two programs (*SI Appendix*, Fig. S3). We manually confirmed all six large (>1 Mb) inversions using dot plots encompassing single, unbroken HiFi contigs (*Materials and Methods*; examples in *SI Appendix*, Fig. S4). We also identified 4,518 segmental duplications (SDs), ranging in size from 1,000 to 4,786,886 bp, using BISER (62), accounting for 8% of the VGP genome (Dataset S6).

The SVs have a mean size of 174 bp, a median size of 338 bp, and encompass 6.5 times more base pairs than do SNPs across the genome (386,804,299 vs. 59,310,087 bp). Most SVs reside in introns and intergenic regions (*SI Appendix*, Fig. S5). SVs overlap more genes than INDELs on average (0.71 vs. 0.54 genes per variant). The largest SV was a 11.3 Mb inversion located in chromosome 1 (Chr1: 86,909,536 to 98,195,717), with the putative centromere located inside the inversion, thus constituting a pericentric inversion (*Materials and Methods*). Over half of the SVs (60%) span repeats as identified by RepeatMasker (63) (Fig. 2; *Materials and Methods*), predominantly simple repeats (16.9%), long terminal repeats (LTRs, 3.2%), long interspersed nuclear elements (LINEs, 2.0%), and unclassified repeats (28.1%). LTRs were particularly enriched in SVs of around 640 bp (Fig. 2*C*) and multiple repeat forms are enriched in SVs of around 6,500 bp (Fig. 2*D*). Nearly 8% (7.7%) SVs span more than one repeat type (Fig. 2*B*). The human-mediated introduction of House Finches to the east (46, 64) has reduced genetic diversity, with fewer SNPs, INDELs, and SVs, lower heterozygosity, and increased ROH compared to the ancestral western population (Fig. 1 *E*–*H* and *SI Appendix*, Results).

Pangenome Gene Graph Captures Genome-Wide Variation in Genic Copy Number. The PGGB pangenome coverage and growth curves indicate a plateau in the discovery of genomic variation (Fig. 3*A*). To examine intraspecific variation in the complement of genes, including genic presence–absence variants (PAV) and copy number variation (CNV), we used Pangene (65) and an annotation of protein-coding genes for the 32 House Finch haplotypes to produce a pangenome gene graph (GFA format) in which each node represents a gene, and an edge between two genes indicates their genomic adjacency on the haplotypes. The annotations were based on the VGP genome using miniprot (66), which aligns the amino acid sequences of proteins to each DNA haplotype assembly. Like the PGGB curves (Fig. 3*A*), the Pangene graph also indicates a plateau in gene discovery (Fig. 3*B*). To reduce false positive gene variation caused by fragmented HiFi contigs, which Pangene filters out and thus flags as potential gene absence, and due to the challenges of mapping protein sequences to genomes (66, 67), we considered confident CNV genes as those with more than two copies in a haplotype, and confident PAV genes as absent in at least five haplotypes (85% filtering threshold, *SI Appendix*, Fig. S6). The resulting graph includes 711

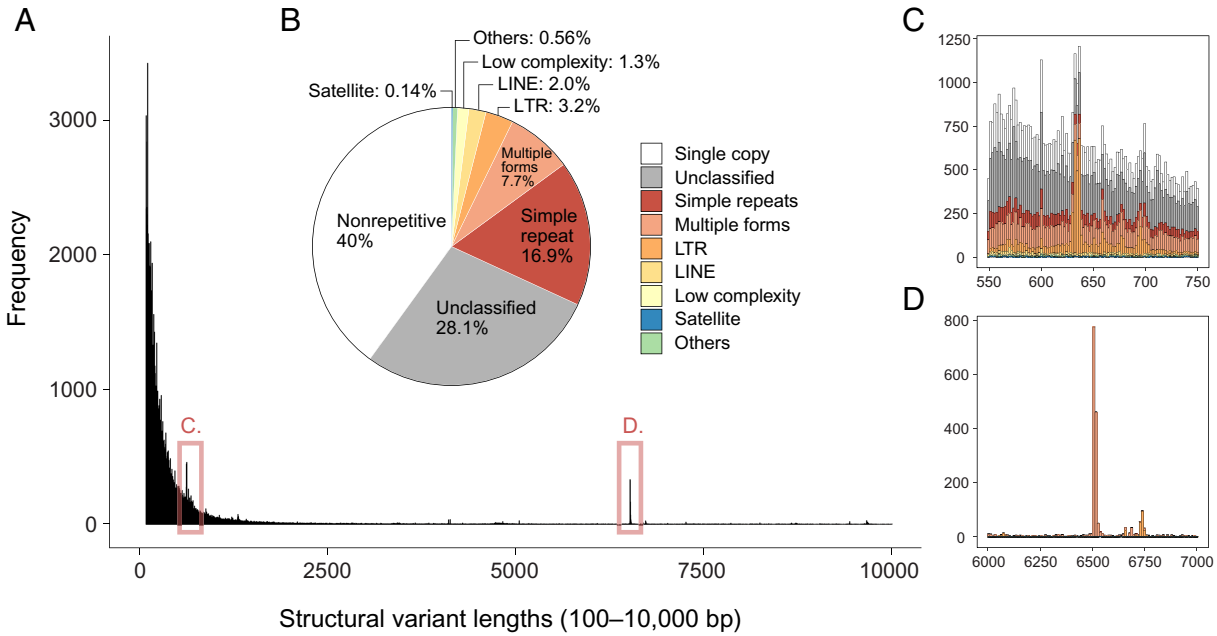

Fig. 2. Most structural variants overlap repeats and transposable elements. (*A*) Histogram illustrating the distribution of lengths of structural variants (SVs), from 100 base pairs (bp) to 10 kilobases (Kb). (*B*) Pie chart delineating SV overlap with various repeat types, transposable elements, and nonrepeats. (*C* and *D*) Detail of histogram sections from panel *A*, displaying SV frequency in the 550 to 750 bp range (*C*) and the 6 to 7 Kb range (*D*).

PAV genes, representing 4.5% of the genes annotated by miniprot (Fig. 3 *C* and *D*), and 180 CNV genes, constituting 1.16% of the annotated genes (Fig. 3*E*). Allelic differentiation (F_{ST}) of PAV genic polymorphisms between the eastern and western populations is low and does not show any convincing outliers (*SI Appendix*, Fig. S7), although slightly higher than that for intergenic SNPs (*SI Appendix*, Fig. S8). These suggest that the distribution of PAV genes may be driven largely by genetic drift.

Fitness Effects of Genome-Wide Structural Variants. We assessed the impact of genome-wide variants on fitness by analyzing their distribution of fitness effects (DFE). We focused on SNPs, INDELs, and SVs in multiple genomic contexts including coding (CDS), noncoding (intron and intergenic), and regulatory (untranslated region, 5′ UTR and 3′ UTR) regions. We estimated the DFE based on the site frequency spectrum (SFS; Fig. 4*A*), accounting for neutral demographic effects, using two maximum likelihood approaches, fastDFE (68) and anavar (69); anavar provides enhanced correction for polarization errors whereas fastDFE offers greater computational efficiency.

The SFSs revealed that all variant types, including SNPs, are predominantly concentrated at low derived allele frequencies, indicating a high incidence of rare variants. SVs and INDELs segregated at much lower average frequencies than SNPs (Fig. 4*A* and *SI Appendix*, Fig. S9). In coding regions, which harbor a higher incidence of rare variants compared to other regions (Fig. 4*A*), most variants are estimated to be deleterious (Fig. 4*B*), ranging from weakly to strongly deleterious as defined by bins of selection coefficients ($\gamma = N_e s$; *Materials and Methods*); the majority of SVs (96%) are estimated to be strongly deleterious. Similarly, in the 3′ and 5′ regulatory UTRs, we also found an enrichment of strongly deleterious variants, whereas INDELs in 3′ and 5′ UTRs tend to be more neutral than in coding regions (Fig. 4*B*). In intronic regions, SNPs and INDELs are estimated to be predominantly neutral, in stark contrast to SVs, which were predominantly identified as strongly deleterious (79%; Fig. 4*B*). Analyses with both fastDFE (Fig. 4*B*) and anavar yielded similar estimates of the DFE (*SI Appendix*, Fig. S10).

We asked whether SV length correlated with the estimated scaled selection coefficient. An analysis of SVs of increasing length, up to 100 bp in 10 bp increments, indicated that larger SVs are generally more deleterious than shorter SVs (*SI Appendix*, Fig. S11). Although SVs are typically defined as >50 bp, we included smaller SVs here to avoid arbitrary cutoffs and capture a broader range of fitness changes. Estimates of the DFE across different repeat types suggest that LTRs and simple repeats exhibit slightly higher deleteriousness compared to other repeat types in intronic regions, whereas in nonintronic regions, the DFE was fairly similar across repeat types (*SI Appendix*, Fig. S12). Further investigation into population-specific fitness effects revealed a slight increase in strongly deleterious SNPs, INDELs, and SVs in the eastern compared to the western population, suggesting that deleterious variants have accumulated detectably in the population with a smaller effective population size (*SI Appendix*, Fig. S13), as predicted by theory (70, 71).

Confirming and Genotyping an 11 Megabase Pericentric Inversion. We were interested in the evolutionary impact and fitness contributions of the largest SV identified, the 11 Mb pericentric inversion. This inversion was confirmed at the HiFi contig level using reference-based alignment (*Materials and Methods*), including dot plots and synteny plots (Fig. 5 *B* and *C* and *SI Appendix*, Fig. S5), visualization of the inversion subgraph using odgi viz (26) (Fig. 5*E*), and analysis of gene arrangements across

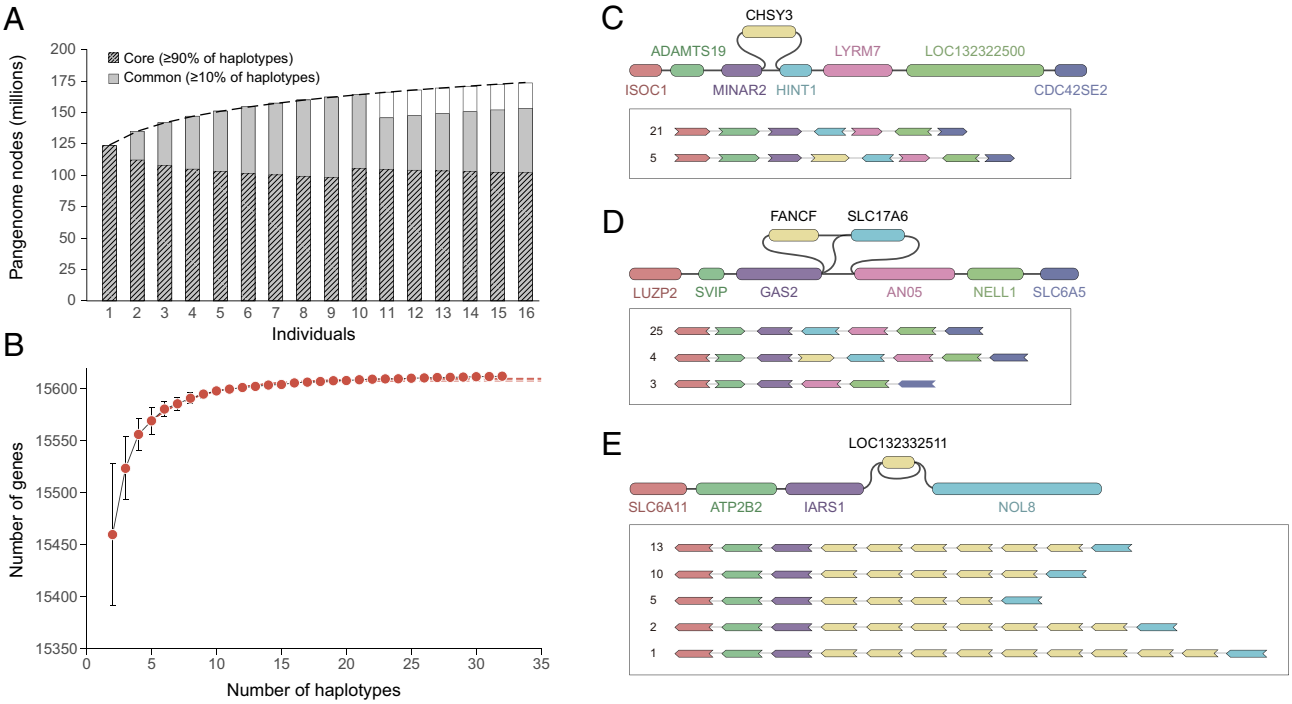

Fig. 3. Pangenome and pangenome gene graph variation. (*A*) PGGB pangenome growth curves show number of new nodes (Y-axis) each sample (X-axis) adds to the graph. Nodes are classified as "core" if present in ≥90% of samples and "common" if present in ≥10%. (*B*) Evaluation of the pangenome gene plateau based on pangene. Data points indicate the average gene count across sampled haplotypes with 100 permutations. The red curve, extrapolating from observed data via a logistic growth model, forecasts the trend. The number of genes obtained from 25, 30, and 32 haplotypes suggest convergence and a plateau in the discovery of new genes. Genes from 32 haplotypes are annotated with miniprot (*Materials and Methods*). (*C*–*E*) Examples of gene presence–absence variation (PAV; *C* and *D*) and copy number variation (CNV; *E*) within the pangenome gene graph constructed using Pangene (*Materials and Methods*). In the graph, a node represents a gene and an edge between two genes indicates their genomic adjacency on the haplotypes. Each panel shows a subgraph of the variant gene, adapted from Bandage visualization (*Top*), and distinct haplotypes (paths) within the subgraph visualized using pangene (*Bottom*).

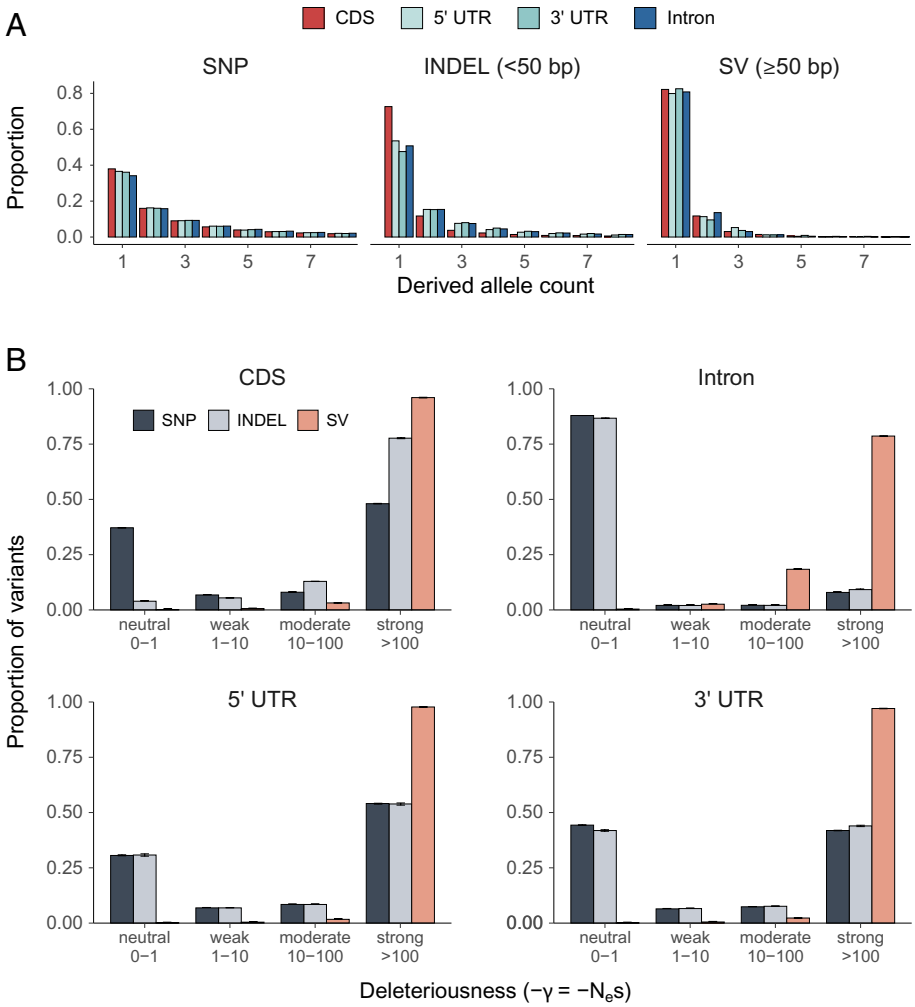

Fig. 4. Fitness effects of genome-wide structural variants and SNPs. (*A*) Unfolded SFS of SNPs, INDELs, and SVs residing in coding sequences (CDS), introns, and 5′ and 3′ UTRs. The figure displays derived allele counts (DAC) ranging from 1 to 8, with DAC 1 to 31 detailed in *SI Appendix*, Fig. S9. (*B*) DFE in bins of population-scaled selection coefficient ($\gamma = N_e s$), reflecting variant deleteriousness, a function of the effective population size (N_e) and the selection coefficient (s). DFE was inferred from SFS of different variant types residing in different genomic categories; a similar pattern was yielded by analyses using both fastDFE (as shown) and anavar (*SI Appendix*, Fig. S10). Error bars indicate 95% CI.

the inversion for haplotypes (Fig. 5*F*; *Materials and Methods*). The large inversion breakpoints were enriched with repeats; one breakpoint occurred near the centromere (gray bars in Fig. 5*A* and interspersed sequences in Fig. 5*C*), where SDs and LTRs were also enriched (Fig. 5*D*). To genotype the inversion in a larger population sample, we leveraged additional datasets consisting of reduced-representation sequencing (RAD seq) data for 108 samples retrieved from Shultz et al. (64); whole-genome short-read sequencing (WGS) data newly generated for 9 samples from central Texas (~15×); and the HiFi data for the 18 pangenome samples (Fig. 6*A*). We genotyped the inversion for these 135 individuals, sampled from across the native and introduced geographic range, including Hawaii, and across different times since encountering the pathogen at the population level (Dataset S1). The RAD-seq dataset included additional outgroup species, including two Cassin's Finches (*Haemorhous cassinii*) and two Purple Finches (*Haemorhous purpureus*).

Principal component analysis (PCA) based on 1,159 genome-wide SNPs called from the dataset clustered the 135 individuals by lineage (western, eastern, Hawaii; Fig. 6*B*; *Materials and Methods*). However, PCA on 20 inversion-associated SNPs formed three distinct clusters with higher heterozygosity in the middle cluster (Fig. 6*C*). These three clusters thus reflect the three inversion karyotypes: the heterokaryotype (Standard [Std]/Inverted [Inv] arrangements), homokaryotype, and alternative homokaryotype (5). The divergence landscape (as measured by d_{XY}) between the inverted and standard haplotypes resembles a "suspension bridge", with low divergence at the center where gene flux (genetic exchange) between arrangements is likely maximal, and high divergence at breakpoints where recombination is likely maximally suppressed, as predicted by coalescent simulations (72, 73) and described in empirical studies (74, 75) (Fig. 6*F*). The inversion haplotype was found in all three species in the genus *Haemorhous* in the studied samples—the House Finch, Cassin's Finch, and Purple Finch—indicating preexisting genetic variation in prior to the origin of House Finches. Although we do not know the ancestral state of the inversion, its minimum age is at least 10 My, which is the time to the most recent common ancestor (TMRCA) of the genus *Haemorhous* (*SI Appendix*, Fig. S14). Additionally, based on the substitution rate of birds (1.9×10^{-9} per site per year) and haplotype divergence (d_{XY}), the inversion's age was estimated at 15 My (Fig. 6*F*; *Materials and Methods*), further supporting the ancient history of this trans-species balanced inversion. Notably, the divergence-based age estimation predates the TMRCA of the samples used in the study (Rosefinch and *Haemorhous*), leaving the ancestral karyotype unresolved.

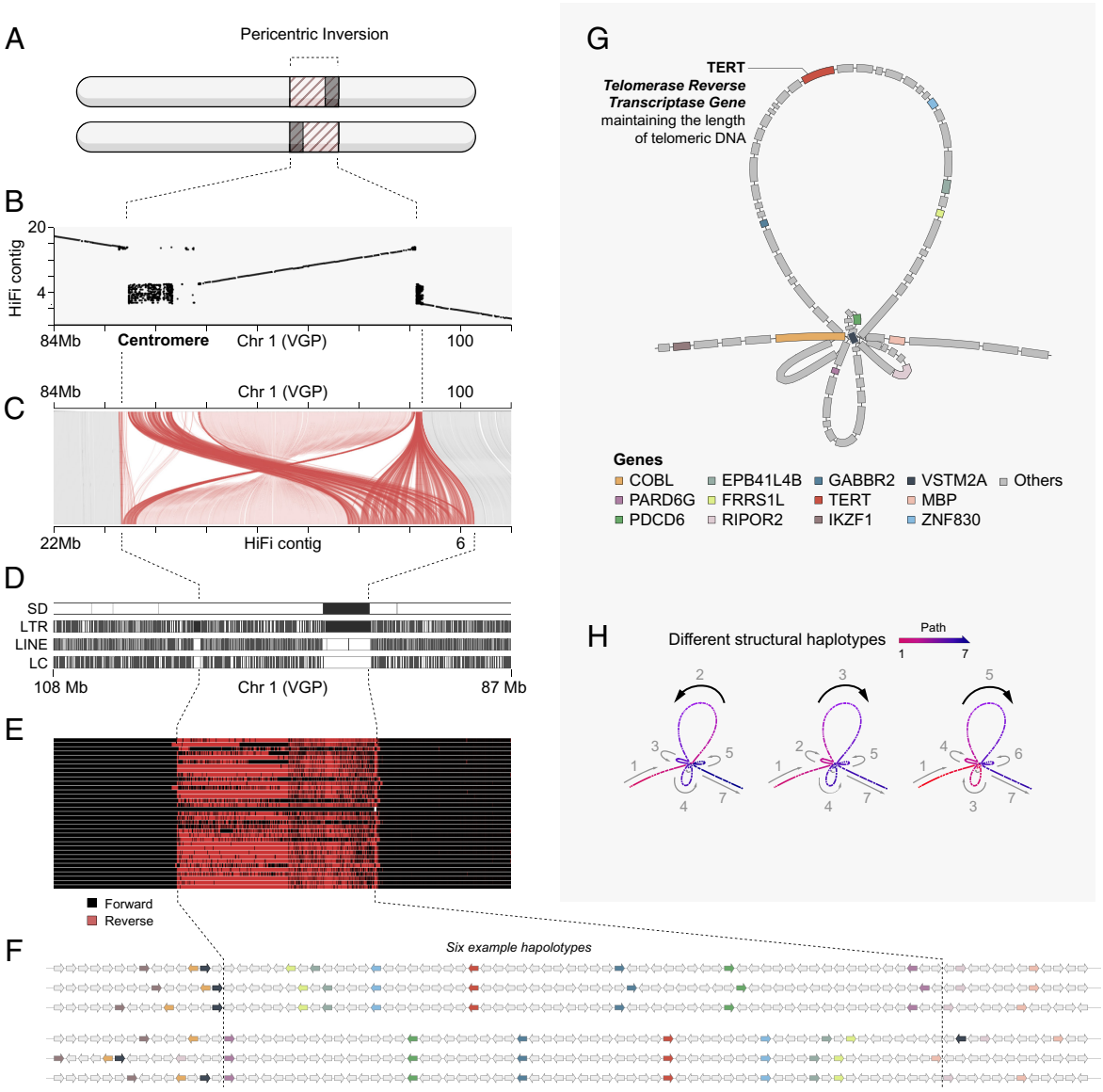

Fig. 5. A pericentric inversion on chromosome 1. (*A*) A schematic of the pericentric inversion involving an inverted centromere (gray bar) and a segment of genetic material from an arm of chromosome 1 (red bar). (*B*) Dot plot of the inversion, displaying the alignment of an NY_2_hap2 haplotype contig ("query", y-axis) with the VGP genome ("reference", x-axis). The inversion spans from 89 to 98 Mb on chromosome 1. The presumed centromere, composed of segmental repeats, is subsumed by the inversion. (*C*) Linear synteny plot of the inversion, correlating with the dot plot (*B*). The interspersed sequences (darker red) represent complex structural variations within and near the centromere. (*D*) SDs and repeat content surrounding the breakpoints. LTR, long terminal repeat; LINE, long interspersed nuclear elements; LC, low complexity. (*E*) Pangenome graph constructed with PGGB and visualized with odgi viz, depicting flips in strand across the inversion. (*F*) Genes within the inversion and nearby breakpoints, with those highlighted in color showing differential expression in response to MG infection, were identified through published experimental infection studies and manual searches for immune function (*Results*). Detailed gene arrangements for all 32 haplotypes are provided in *SI Appendix*, Fig. S17. The key to these genes is shown in (*G*), where a pangenome gene graph displays the arrangement of genes within and surrounding the inversion. (*H*) Examples of haplotype paths among the 32 haplotypes, indicating different structural haplotypes around the inversion and its breakpoints.

Heterokaryotypes for the Large Inversion Have Increased in Frequency during the *Mycoplasma* Epizootic. We found that the inversion genotype frequency shifted significantly between pre- and postepizootic birds. Specifically, the frequency of the heterokaryotype increased by 62.5%, rising from 0.4 to 0.65 (Fig. 6*D*) across pooled pre- and postepizootic populations. This increase was evident in both eastern and western populations (*SI Appendix*, Fig. S15). Over the same time period, the frequency of both homokaryotypes (Std/Std and Inv/Inv) decreased, with one form declining by 90%, from 0.2 to 0.02 (Fig. 6*D*). Eleven years postepizootic, the heterokaryotype genotype reached a frequency of 1 across pooled populations (Fig. 6*E*). Additionally, the inversion exhibited some of the highest population differentiation as measured by F_{ST} between individuals with two different homokaryotypes (i.e., estimates of F_{AT}) (76) (*SI Appendix*, Fig. S16). However, we found that population differentiation of the inversion before and after the epizootic is lower than other genomic regions (*SI Appendix*, Fig. S16), consistent with balancing selection diminishing population differentiation (77). These results suggest that the inversion confers a fitness advantage, because heterokaryotypes are maintained as a balanced polymorphism within the population, rather than one or the other inversion allele becoming fixed by drift or selection.

We investigated the genic content of the inversion and the diversity of inversion-associated haplotypes in the pangenome graph. Two recent studies, by Veetil et al. (78) and Henschen et al. (49), identified ca. 1,780 genes differentially expressed in immune responses of House Finches to MG using experimental infections,

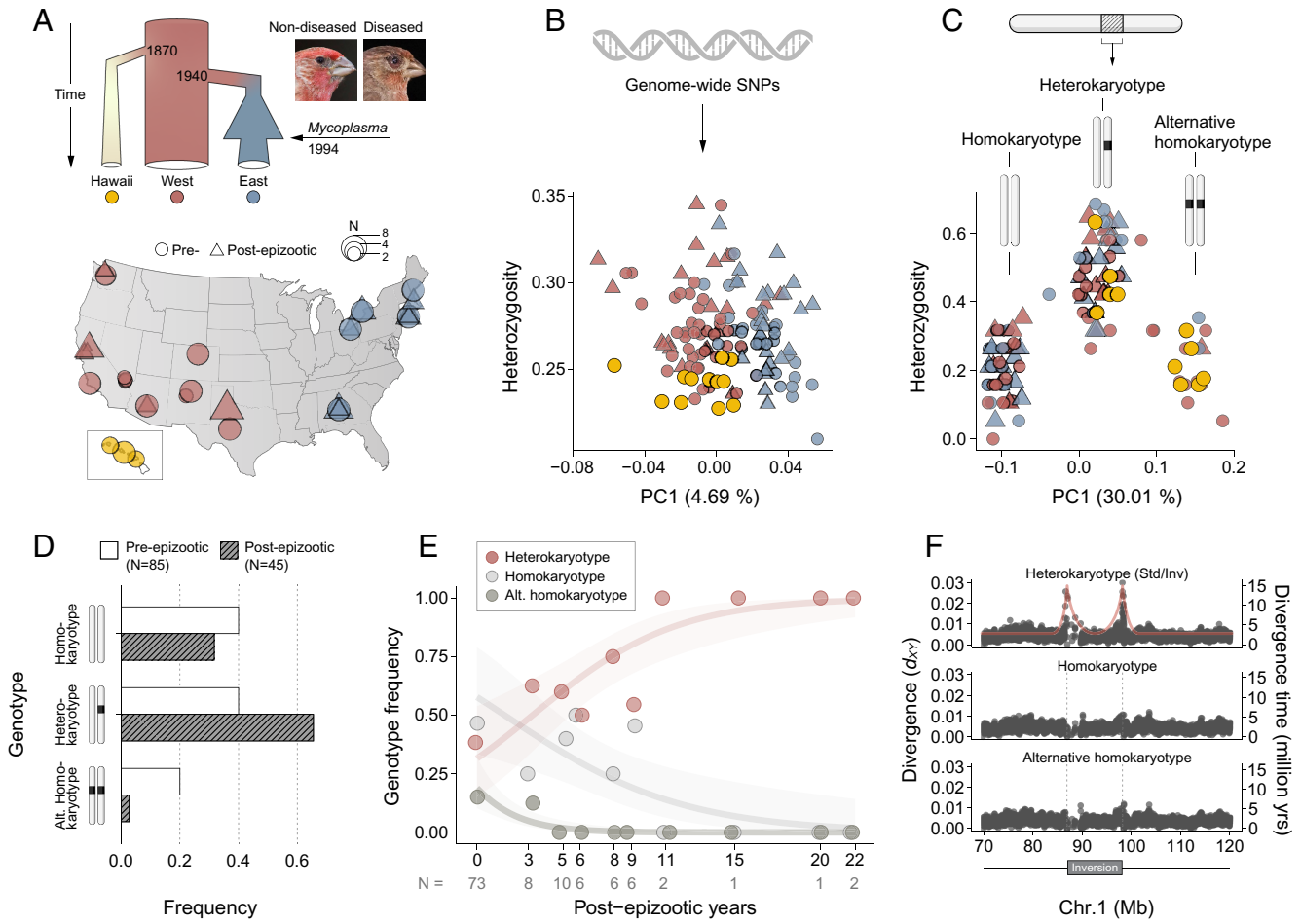

Fig. 6. Association between inversion genotype frequency and time since pathogen exposure. (*A*) Demographic history and map of genetic sampling of House Finch populations, based on data from HiFi sequencing in this study, Shultz et al. [(64); RAD sequencing], and newly sequenced samples (WGS; see *Materials and Methods*). Demographic plot adapted from Shultz et al. (64); bird photos retrieved from Bonneaud et al. (42). (*B*) PCA and individual heterozygosity based on genome-wide SNPs. The X-axis shows Principal Component 1 (PC1), accounting for the most variance. (*C*) PCA and individual heterozygosity based on SNPs within the inversion, indicating three inversion karyotypes: heterokaryotype, homokaryotype, and alternative homokaryotype. (*D*) Inversion genotype shifts pre- and postepizootic, suggesting a heterozygote advantage postepizootic. (*E*) Genotype (karyotype) frequency of the inversion shifts over the course of the epizootic. Birds from Hawaii, with only pre-epizootic data, are not considered in *D* and *E*. (*F*) Divergence time of inversion karyotypes inferred from genetic distance (d_{XY}) between the two haplotypes per 10Kb sliding windows and substitution rate of birds 1.9×10^{-9} substitutions per site per year (*Materials and Methods*). Dots present divergence time per sliding window within and outside the inversion.

suggesting a number of candidates for involvement in the inversion. Of the infection-responsive genes, five (GABBR2, ZNF830, EPB41L4B, FRRS1L, and PDCD6) were found to reside in the inversion, and five (IKZF1, PARD6G, VSTM2A, COBL, and MBP) were found near the inversion breakpoints. Additionally, by searching the NCBI database, we found two immune-related genes (RIPOR2 and RIPK2) near the breakpoints, and in the middle of the inversion a telomerase reverse transcriptase gene (TERT), which plays a crucial role in the maintenance and lengthening of telomeres (Fig. 5 *F* and *G* and *SI Appendix*, Fig. S17). The discovery of the TERT gene within the inversion prompted our interest in variation in telomere length in the context of the epizootic, and we found telomere length decreases with longer times of population exposure to mycoplasmal pathogen (*SI Appendix*, Fig. S18 and *Results and Discussion*). The functions of above-mentioned genes and references are listed in the Dataset S7. To examine the positional diversity of these genes across haplotypes spanning the inversion, we examine the pangenome gene graph, which confirmed the inversion and identified different haplotype paths through it (Fig. 5 *F–H*). These results indicate dynamic rearrangements of synteny and genes with functions in immunity and telomerase activity in the inversion and surrounding its breakpoints, potentially contributing to adaptive disease resistance.

Discussion

Using long-read sequencing and pangenome approaches on a panel of House Finches, we estimated the fitness consequences of genome-wide variants, including SVs, which have been underrepresented in short-read sequencing data. We detected the imprint of the human-induced introduction of House Finches into the eastern US, including a reduction of genetic diversity for both SNPs and SVs, as well as increased signatures of inbreeding in the introduced eastern population (Fig. 1 *E–H*). We also found lower telomere motif abundance in individuals from populations that had been exposed to the *Mycoplasma* pathogen for longer periods of time. Finally, we detected an association between the largest SV, an 11 Mb pericentric inversion, and time since population exposure to the pathogen, with a signature of balancing selection in inversion genotype (karyotype) frequencies.

Our genome-wide assessment of variants revealed that the vast majority of SVs are estimated to be strongly deleterious across the genome, more so than smaller variants like SNPs and INDELs, and that variants in coding and regulatory regions are more deleterious than those in noncoding regions. It is important to note that the statistical models we used only allowed for the estimation of negative or neutral selection coefficients, based on the

assumption that beneficial mutations are typically rare and subject to strong positive selection, leading to their rapid fixation in the population and minimal contribution to DFE (71, 79). Still, these methods can distinguish between neutral or close-to-neutral and deleterious variants, depending on the departure of the SFS from the putatively neutral SFS (Materials & Methods). The trend of estimated strongly negative selection against SVs is observed across coding, noncoding, and regulatory sequences such as 3′ and 5′ UTRs. Different studies suggest that the fitness effects of SVs arise from both direct and indirect mechanisms. Direct effects include disruption of gene function by altering coding regions and regulatory elements, or by inducing position effects, changing gene dosage, or influencing gene expression at breakpoints (2, 73, 80, 81). Indirectly, SVs can affect fitness by suppressing recombination (82, 83). Our findings also support the hypothesis that the length of SVs relates to their estimated harmfulness (73, 84). Smaller INDELs were generally estimated to be more neutral in noncoding regions compared to larger SVs (Fig. 4*B*), and we observed an increasing trend in deleteriousness with SV size (*SI Appendix*, Fig. S11). Additionally, we found that larger variants tend to overlap more genes than smaller ones. Given the outlined mechanisms, larger variants likely exert a more pronounced impact on fitness by disrupting gene functions and/or capturing large genomic segments, increasing the likelihood of trapping additional deleterious variants within the segment.

Empirical studies suggest that a population bottleneck could impact deleterious variation in complex ways, either increasing or reducing the genetic load (71). We found tentative evidence that the eastern population is enriched with more strongly deleterious variants of all types than the western population, suggesting that even a slightly lower effective population size could weaken the efficacy of purifying selection, thus increasing the accumulation of deleterious mutations (85–87). Our findings highlight the necessity of including SVs in studies of genetic diversity and fitness effects to fully understand the evolutionary dynamics of bottlenecks and other demographic events. Furthermore, our estimates of the DFE indicate that LTRs and simple repeats exhibit slightly higher deleteriousness in intronic regions compared to other repeat types, such as LINEs and low-complexity repeats, even though simple repeats might sometimes provide adaptive variation that enhances fitness (88–90). This elevated deleteriousness may be attributed to the intrinsic properties of these repeat types: they are known to contribute to genomic instability and disrupt gene regulation. LTRs, a category of retrotransposons, may alter gene expression or cause mutations through transposition events (91), whereas the structure of simple repeats makes them prone to replication slippage, leading to insertions or deletions that can interfere with gene function (92). The presence of these repeats in intronic regions might also increase the likelihood of splicing errors or misregulation, further contributing to their deleterious effects.

Beyond the overall detrimental effects of SVs, our study improves understanding of the potential roles and mechanisms of SV in adaptation. We highlighted a large long-term trans-species balanced inversion whose genotype frequencies correlated strongly with population time to exposure to the *Mycoplasma* pathogen, and therefore potentially associated with adaptive evolution in response to disease. This inversion polymorphism is found across all three *Haemorhous* species, with an estimated minimum age of 10 Mya, pointing to the role of standing genetic variation in adapting to environmental challenges, such as immune response (93) and pathogen resistance (15). We considered the possibility that the distribution of the inversion within and among species could be due to recurrent mutation, as has been found in some inversions in humans (94). However, the high level of divergence between different haplotypes across the inversion suggests instead that it is very old. The substantial genotype frequency changes imply that selection is at play. Positive selection may establish an inversion through indirect (linked) selection—such as when the inversion captures a locally adaptive haplotype and "hitchhikes" with it—or through direct positive selection, as when an advantageous mutation occurs near the breakpoints (5, 6, 73, 75, 95–98). The suspension bridge divergence landscape between the noninverted and inverted karyotypes (Fig. 6*F*) suggests either neutral processes or selection acting directly on the breakpoints, according to coalescent models (72) and empirical observations (74, 75, 99, 100). However, given that old neutral inversions are likely to be either fixed or lost (73), selection is a more plausible explanation in this case. Additionally, we identified genes near breakpoints potentially related to the House Finch immune response (Fig. 5 *F* and *G* and *SI Appendix*, Fig. S17), supporting the "adaptive breakpoints" scenario (95). Overall, our results tentatively suggest a role of SVs in influencing the expression of adjacent genes, which could decrease or increase an organism's fitness depending on gene functions.

Variable structural rearrangements encompassing genes, including the TERT gene, were identified within and near to the inversion across the 32 haplotype assemblies. Such rearrangements may impact telomerase and immune functions and suggest a putative position effect on genes within and surrounding the inversion due to shifts in their genomic location or surrounding chromatin environment. Additionally, the discovery that large inversion breakpoints occurred near the centromere and were enriched with SDs and LTRs is consistent with the hypothesis that the accumulation of transposable elements in pericentromeric regions may drive SD and inversion formation (101). However, it is important to note that our study did not investigate the mechanisms underlying the balanced polymorphisms of the 11 Mb inversion. Potential factors contributing to this balance could include epistatic balancing selection and negative frequency-dependent selection (4, 6, 15, 73, 97, 98, 102). To gain a deeper understanding of the genetic architecture of the inversion and the mechanisms of its balancing selection, further increasing marker density at the population level, such as through WGS data, and experimental studies, are necessary.

Large inversions have been associated with avian morphs, behavior, and mating strategies (103). The largest 11.3 Mb inversion we identified is comparable in size to functional inversions in other avian systems, such as a 7.4 Mb inversion in Chickens (*Gallus gallus*) and a 4.5 Mb inversion in Ruffs (*Philomachus pugnax*) (11, 104). Larger inversions have been also observed in other birds, including a 55 Mb inversion in Redpoll Finches (*Acanthis* spp.) (105), a 115 Mb inversion in Quail (*Coturnix coturnix*) (106), and a >100 Mb inversion in White-throated Sparrows (12). However, birds generally exhibit fewer and smaller inversions compared to mammals such as humans and the deer mouse (*Peromyscus maniculatus*). The human genome contains over 1,000 inversions, including >100 larger than 1Mb (94, 107), whereas deer mice have 21 inversions larger than 1Mb and potentially thousands of smaller ones in its genome (7, 108). This suggests that avian inversions are generally shorter and less frequent than in mammals, possibly due to their more streamlined genomes (109), which contain fewer repeats and transposable elements driving the formation of inversions (103, 108).

Certain limitations must be acknowledged to fully interpret our findings. First, the relatively small sample size (16 individuals) cannot fully capture the diversity of the pangenome for all House Finches. Although our evaluation of pangenome graphs suggests a near plateau in the discovery of new core variation and genes

using 32 haplotypes, we have potentially overlooked rare genetic variants. Nevertheless, previous studies indicate that more than 8 haplotypes/alleles from a randomly mating population are sufficient to have a high probability of capturing the root of the coalescent tree, which is a major determinant of genetic diversity (110, 111). Second, using an outgroup that diverged 12.9 Mya could affect our ability to accurately deduce ancestral states and derived allele frequencies, which are the basis for assessing fitness effects. To mitigate this, we have employed rigorous criteria for polarizing sites to only homologous outgroup genotypes (*Materials and Methods*) and using anavar, which models and accounts for errors in ancestral state polarization (69). These measures enhance the reliability of our conclusions, despite the challenges of relying on outgroup-based ancestral state reconstructions.

This study underscores the use of combining population-scale long-read sequencing with pangenomic methods in molecular ecology. This approach offers a more complete exploration of genetic variants than is possible with short-read data alone. Other potential approaches to cataloging SVs in natural populations include mapping short reads from many individuals to a pangenome graph composed of a few individuals (112). This approach could be a favorable economic alternative to sequencing all individuals with long-read methods as we have done here. Looking ahead, the fusion of long-read and pangenomic approaches promises to unveil novel findings about genetic diversity and evolutionary adaptation.

Materials and Methods

Sampling and Sequencing. We retrieved 16 House Finch individuals from the western (CA, WA, AZ, NM) and eastern US (NY, MA, OH, AL) and a Common Rosefinch as the outgroup (Dataset S1), accessioned in the Museum of Comparative Zoology (MCZ) at Harvard. High-molecular-weight DNA was extracted using the Qiagen MagAttract HMW DNA Kit and assessed with TapeStation, Qubit, and NanoDrop at Harvard's Bauer Core Facility. PacBio HiFi sequencing was conducted at the University of Delaware. Exposure time to *Mycoplasma* was estimated from the collection year minus the earliest documented arrival of the pathogen. For population genomic analyses, RAD-seq data were obtained from 104 House Finches and other related species from Shultz et al. (64); nine House Finch samples were whole-genome sequenced at ~15× depth on an Illumina NovaSeq at Harvard Bauer Core Facility (*SI Appendix*, *Methods*).

Genome Assembly and Annotation. We generated two de novo haplotype genome assemblies per sample using Hifiasm (59) and assessed them with assembly-stats and BUSCO scores (Dataset S2). The VGP genome was annotated using RNA-seq data, homology information, ab initio gene prediction, and projection-based gene prediction. Repetitive elements and SDs were identified with RepeatMasker (63) and BISER (62), respectively (*SI Appendix*, *Methods*).

Pangenome Graph Construction and Variant Decomposition. PGGB pipeline was used to construct pangenome graphs for each chromosome. Variants were called from the graphs using tools vg (113), vcfwave (114), vcflib (114), and bcftools (115). Variants were classified based on allele sizes: SNPs if both reference (REF) and alternative (ALT) alleles were 1 bp; insertions (INS) and deletions (DEL) for size differences under 50 bp; and as SVs (SV-INS and SV-DEL) for differences over 50 bp. Variant polarization was based on the ancestral allele (outgroup). Complex types and multiallelic variants were also categorized (*SI Appendix*, *Methods*).

Identifying and Genotyping Inversions. To enhance inversion discovery, we additionally performed SVIM-asm (60) and SyRi (61), acknowledging the difficulty of accurately identifying all inversions using a single tool (116). Inversions were confirmed using dot and synteny plots based on pairwise alignments of haplotypes against the VGP genome using custom R scripts. Verified inversions had two breakpoints within at least one HiFi contig. The 11.3 Mb inversion was visualized using odgi (26). Inversion genotyping across 135 samples involved mapping sequencing reads (RAD-seq, WGS, HiFi) to the VGP genome, with PCA on SNPs performed using SNPRelate (117). The inversion age was estimated by dividing genetic distance (d_{XY}) by substitution rates for birds (118) in a 10 Kb sliding window across the inversion (*SI Appendix*, *Methods*).

Estimating the DFEs. DFEs was estimated using fastDFE (68) and anavar (69) which use the SFS to estimate the population-scaled mutation rate ($\theta = 4N_e\mu$; N_e is the effective population size and μ is the per site per generation mutation rate) and shape and scale parameters for a gamma distribution of population-scaled selection coefficients ($\gamma = 4N_e s$; s is the selection coefficient). DFEs were fitted using intergenic sites as the neutral reference to control for demographic and polarization errors. We present the gamma distributions as the proportion of variants falling into four bins of selection coefficients (γ) representing the scaled selection coefficients of variants: neutral ($0 \leq -N_e s \leq 1$), weak ($1 < -N_e s \leq 10$), moderate ($10 < -N_e s \leq 100$), and strong ($-N_e s > 100$), and derived 95% CI through bootstrapping in fastDFE and permutation in anavar (*SI Appendix*, *Methods*).

Analysis of Genetic Diversity. Runs of homozygosity were identified using PLINK (119) with SNPs from the PGGB VCF. Individual heterozygosity for SNPs, insertions, and deletions was calculated in regions identified as confident by Dipcall (120) (*SI Appendix*, *Methods*).

Data, Materials, and Software Availability. The assemblies, raw HiFi, Iso-Seq, and whole-genome sequencing (WGS) data will be made accessible from the NCBI (PRJNA1101522) (121). Analytical scripts are currently available on GitHub at https://github.com/fangbohao (122). PGGB graphs for each chromosome, along with variant data, can be accessed on Dyrad (https://doi.org/10.5061/dryad.hhmgqnkqb) (123).

ACKNOWLEDGMENTS. We thank Geoffrey Hill (Auburn University), Allison Shultz, Jonathan Schmitt, Flavia Termignoni Garcia, and Kathrin Näpflin for their contributions to the House Finch collections at the Museum of Comparative Zoology, Harvard University. We are grateful to Amberleigh Henschen (University of Memphis) and the VGP for the collection and assembly of the VGP House Finch genome. We also thank Timothy Sackton, Heng Li, Erik Garrison, Andrea Guarracino, and Danielle Khost for technical advice, and Gabriel David and Jacob Höglund for valuable discussions. We acknowledge the FASRC Cannon cluster at Harvard University for computational resources and the reviewers for their constructive comments. This work was funded by Harvard University, the Harvard Global Institute, Harvard China Funds (to S.V.E. and B.F.), and the Finnish Cultural Foundation, Grant No. 00211290 (to B.F.).

1. E. J. Hollox, L. W. Zuccherato, S. Tucci, Genome structural variation in human evolution. *Trends Genet.* **38**, 45–58 (2021), 10.1016/j.tig.2021.06.015.
2. C. Mérot, R. A. Oomen, A. Tigano, M. Wellenreuther, A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol. Evol.* **35**, 561–572 (2020).
3. S. S. Ho, A. E. Urban, R. E. Mills, Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
4. M. Wellenreuther, L. Bernatchez, Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).
5. C. Merot, Making the most of population genomic data to understand the importance of chromosomal inversions for adaptation and speciation. *Mol. Ecol.* **29**, 2513–2516 (2020).
6. A. A. Hoffmann, C. M. Sgro, A. R. Weeks, Chromosomal inversion polymorphisms and adaptation. *Trends Ecol. Evol.* **19**, 482–488 (2004).
7. E. R. Hager *et al.*, A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. *Science* **377**, 399–405 (2022).
8. F. C. Jones *et al.*, The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
9. Z. Zhang *et al.*, Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *Plant Cell* **27**, 1595–1604 (2015).
10. K. E. Delmore *et al.*, Structural genomic variation and migratory behavior in a wild songbird. *Evol. Lett.* **7**, 401–412 (2023).
11. S. Lamichhaney *et al.*, Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nat. Genet.* **48**, 84–88 (2016).
12. E. M. Tuttle *et al.*, Divergence and functional degradation of a sex chromosome-like supergene. *Curr. Biol.* **26**, 344–350 (2016).
13. E. M. Leffler *et al.*, Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* **356**, eaam6393 (2017).
14. J. Luo *et al.*, Genome-wide copy number variant analysis in inbred chickens lines with different susceptibility to Marek's disease. *G3 (Bethesda)* **3**, 217–223 (2013).

15. A. A. Hardikar, B. B. Nath, Chromosomal polymorphism is associated with nematode parasitism in a natural population of a tropical midge. *Chromosoma* **110**, 58–64 (2001).
16. A. Dolatabadian *et al.*, Characterization of disease resistance genes in the *Brassica napus* pangenome reveals significant structural variation. *Plant Biotechnol. J.* **18**, 969–982 (2020).
17. W. De Coster, M. H. Weissensteiner, F. J. Sedlazeck, Towards population-scale long-read sequencing. *Nat. Rev. Genet.* **22**, 572–587 (2021), 10.1038/s41576-021-00367-3.
18. M. U. Ahsan, Q. Liu, J. E. Perdomo, L. Fang, K. Wang, A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nat. Methods* **20**, 1143–1158 (2023), 10.1038/s41592-023-01932-w.
19. X. Duan, M. Pan, S. Fan, Comprehensive evaluation of structural variant genotyping methods based on long-read sequencing data. *BMC Genomics* **23**, 324 (2022).
20. J. M. Eizenga *et al.*, Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
21. W.-W. Liao *et al.*, A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
22. R. M. Sherman, S. L. Salzberg, Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
23. H. Li, X. Feng, C. Chu, The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
24. F. Andreace, P. Lechat, Y. Dufresne, R. Chikhi, Construction and representation of human pangenome graphs. *Genome Biol.* **24**, 274 (2023), 10.1101/2023.06.02.542089.
25. E. Garrison *et al.*, Building pangenome graphs. *Nat. Methods*, 10.1038/s41592-024-02430-3 (2024).
26. A. Guarracino, S. Heumos, S. Nahnsen, P. Prins, E. Garrison, ODGI: Understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).
27. F. Andreace, P. Lechat, Y. Dufresne, R. Chikhi, Comparing methods for constructing and representing human pangenome graphs. *Genome Biol.* **24**, 274 (2023).
28. D. Porubsky, E. E. Eichler, A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell* **187**, 1024–1037 (2024), 10.1016/j.cell.2024.01.002.
29. Y. Gao *et al.*, A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
30. T. Wang *et al.*, The Human Pangenome Project: A global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
31. M. Schreiber, M. Jayakodi, N. Stein, M. Mascher, Plant pangenomes for crop improvement, biodiversity and evolution. *Nat. Rev. Genet.* **25**, 563–577 (2024), 10.1038/s41576-024-00691-4.
32. K. Wang *et al.*, Duck pan-genome reveals two transposon insertions caused bodyweight enlarging and white plumage phenotype formation during evolution. *iMeta* **3**, e154 (2023).
33. R. Li *et al.*, A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Res.* **33**, 463–477 (2023), 10.1101/gr.277372.122.
34. Y. Zhou *et al.*, Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. *Genome Res.* **32**, 1585–1601 (2022), 10.1101/gr.276550.122.
35. A. S. Leonard *et al.*, Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat. Commun.* **13**, 3012 (2022).
36. E. S. Rice *et al.*, A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. *BMC Biol.* **21**, 267 (2023).
37. H. Tettelin, D. Riley, C. Cattuto, D. Medini, Comparative genomics: The bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472–477 (2008).
38. J. Liao *et al.*, Nationwide genomic atlas of soil-dwelling *Listeria* reveals effects of selection and population ecology on pangenome evolution. *Nat. Microbiol.* **6**, 1021–1030 (2021).
39. F. X. Quah *et al.*, A pangenomic perspective of the Lake Malawi cichlid radiation reveals extensive structural variation driven by transposable elements. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.03.28.587230 (Accessed 31 April 2024).
40. S. Secomandi *et al.*, Pangenomics provides insights into the role of synanthropy in barn swallow evolution. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.03.28.486082 (Accessed 1 April 2022).
41. Z. Wang, K. Farmer, G. E. Hill, S. V. Edwards, A cDNA macroarray approach to parasite-induced gene expression changes in a songbird host: Genetic response of house finches to experimental infection by *Mycoplasma gallisepticum*. *Mol. Ecol.* **15**, 1263–1273 (2006).
42. C. Bonneaud *et al.*, Rapid evolution of disease resistance is accompanied by functional changes in gene expression in a wild bird. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7866–7871 (2011).
43. C. Bonneaud *et al.*, Rapid antagonistic coevolution in an emerging pathogen and its vertebrate host. *Curr. Biol.* **28**, 2978–2983.e2975 (2018).
44. N. Backstrom, D. Shipilina, M. P. Blom, S. V. Edwards, Cis-regulatory sequence variation and association with *Mycoplasma* load in natural populations of the house finch (*Carpodacus mexicanus*). *Ecol. Evol.* **3**, 655–666 (2013).
45. J. C. Owen, D. M. Hawley, K. P. Huyvaert, *Infectious Disease Ecology of Wild Birds* (Oxford University Press, 2021).
46. D. M. Hawley, D. Hanley, A. A. Dhondt, I. J. Lovette, Molecular evidence for a founder effect in invasive house finch (*Carpodacus mexicanus*) populations experiencing an emergent disease epidemic. *Mol. Ecol.* **15**, 263–275 (2006).
47. Z. Wang, A. J. Baker, G. E. Hill, S. V. Edwards, Reconciling actual and inferred population histories in the house finch (*Carpodacus mexicanus*) by AFLP analysis. *Evolution* **57**, 2852–2864 (2003).
48. W. M. Hochachka, A. A. Dhondt, Density-dependent decline of host abundance resulting from a new infectious disease. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5303–5306 (2000).
49. A. E. Henschen *et al.*, Rapid adaptation to a novel pathogen through disease tolerance in a wild songbird. *PLoS Pathog.* **19**, e1011408 (2023).
50. C. Bonneaud, S. L. Balenger, J. Zhang, S. V. Edwards, G. E. Hill, Innate immunity and the evolution of resistance to an emerging infectious disease in a wild bird. *Mol. Ecol.* **21**, 2628–2639 (2012).
51. K. L. Farmer, G. E. Hill, S. R. Roberts, Susceptibility of wild songbirds to the house finch strain of *Mycoplasma gallisepticum*. *J. Wildl. Dis.* **41**, 317–325 (2005).
52. B. K. Hartup, G. V. Kollias, Field investigation of *Mycoplasma gallisepticum* infections in house finch (*Carpodacus mexicanus*) eggs and nestlings. *Avian Dis.* **43**, 572–576 (1999).
53. A. A. Dhondt, K. V. Dhondt, W. M. Hochachka, D. H. Ley, D. M. Hawley, Response of house finches recovered from *Mycoplasma gallisepticum* to reinfection with a heterologous strain. *Avian Dis.* **61**, 437–441 (2017).
54. C. R. Faustino *et al.*, *Mycoplasma gallisepticum* infection dynamics in a house finch population: Seasonal variation in survival, encounter and transmission rate. *J. Anim. Ecol.* **73**, 651–669 (2004).
55. W. M. Hochachka, A. P. Dobson, D. M. Hawley, A. A. Dhondt, Host population dynamics in the face of an evolving pathogen. *J. Anim. Ecol.* **90**, 1480–1491 (2021).
56. T. D. Price *et al.*, Niche filling slows the diversification of Himalayan songbirds. *Nature* **509**, 222–225 (2014).
57. D. M. Hooper, T. D. Price, Chromosomal inversion differences correlate with range overlap in passerine birds. *Nat. Ecol. Evol.* **1**, 1526–1534 (2017).
58. J. J. Elliott, R. S. Arbib Jr., Origin and status of the house finch in the eastern United States. *Auk* **1**, 31–37 (1953).
59. H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
60. D. Heller, M. Vingron, SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2020).
61. M. Goel, H. Sun, W. B. Jiao, K. Schneeberger, SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
62. H. Iseric, C. Alkan, F. Hach, I. Numanagic, Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol. Biol.* **17**, 4 (2022).
63. J. M. Flynn *et al.*, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9451–9457 (2020).
64. A. J. Shultz, A. J. Baker, G. E. Hill, P. M. Nolan, S. V. Edwards, SNPs across time and space: Population genomic signatures of founder events and epizootics in the House Finch (*Haemorhous mexicanus*). *Ecol. Evol.* **6**, 7475–7489 (2016).
65. H. Li, M. Marin, M. R. Farhat, Exploring gene content with pangenome gene graphs. *Bioinformatics*, 10.1093/bioinformatics/btae456 (2024).
66. H. Li, Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
67. H. Li, R. Durbin, Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.* **25**, 658–670 (2024), 10.1038/s41576-024-00718-w.
68. J. Sendrowski, T. Bataillon, fastDFE: Fast and flexible joint inference of the distribution of fitness effects. *Mol. Biol. Evol.* **41**, msae070 (2024), 10.1101/2023.12.04.569837.
69. H. J. Barton, K. Zeng, New methods for inferring the distribution of fitness effects for INDELs and SNPs. *Mol. Biol. Evol.* **35**, 1536–1546 (2018).
70. T. Ohta, The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286 (1992).
71. J. Robinson, C. C. Kyriazis, S. C. Yuan, K. E. Lohmueller, Deleterious variation in natural populations and implications for conservation genetics. *Annu. Rev. Anim. Biosci.* **11**, 93–114 (2023).
72. R. F. Guerrero, F. Rousset, M. Kirkpatrick, Coalescent patterns for chromosomal inversions in divergent populations. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **367**, 430–438 (2012).
73. E. L. Berdan *et al.*, How chromosomal inversions reorient the evolutionary process. *J. Evol. Biol.* **36**, 1761–1782 (2023), 10.1111/jeb.14242.
74. C. Cheng *et al.*, Ecological genomics of *Anopheles gambiae* along a latitudinal cline: A population-resequencing approach. *Genetics* **190**, 1417–1432 (2012).
75. M. Kapun, T. Flatt, The adaptive significance of chromosomal inversion polymorphisms in *Drosophila melanogaster*. *Mol. Ecol.* **28**, 1263–1282 (2019).
76. B. Charlesworth, M. Nordborg, D. Charlesworth, The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* **70**, 155–174 (1997).
77. D. Y. C. Brandt, J. Cesar, J. Goudet, D. Meyer, The effect of balancing selection on population differentiation: A study with HLA genes. *G3 (Bethesda)* **8**, 2805–2815 (2018).
78. N. Kuttiyarthu Veetil *et al.*, Varying conjunctival immune response adaptations of house finch populations to a rapidly evolving bacterial pathogen. *Front. Immunol.* **15**, 1250818 (2024).
79. T. R. Booker, Inferring parameters of the distribution of fitness effects of new mutations when beneficial mutations are strongly advantageous and rare. *G3* **10**, 2317–2326 (2020).
80. M. Kirkpatrick, N. Barton, Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
81. N. B. Stewart, R. L. Rogers, Chromosomal rearrangements as a source of new gene formation in *Drosophila yakuba*. *PLoS Genet.* **15**, e1008314 (2019).
82. K. N. Crown, D. E. Miller, J. Sekelsky, R. S. Hawley, Local inversion heterozygosity alters recombination throughout the genome. *Biol.* **28**, 2984–2990.e2983 (2018).
83. M. Todesco *et al.*, Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).
84. T. Connallon, C. Olito, Natural selection and the distribution of chromosomal inversion lengths. *Mol. Ecol.* **31**, 3627–3641 (2021).
85. G. Bertorelle *et al.*, Genetic load: Genomic estimates and applications in non-model animals. *Nat. Rev. Genet.* **23**, 492–503 (2022).
86. T. Leroy *et al.*, Island songbirds as windows into evolution in small populations. *Curr. Biol.* **31**, 1303–1310.e1304 (2021).
87. A. Eyre-Walker, P. D. Keightley, The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
88. R. Gemayel, M. D. Vinces, M. Legendre, K. J. Verstrepen, Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–477 (2010).
89. R. Moxon, C. Bayliss, D. Hood, Bacterial contingency loci: The role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* **40**, 307–333 (2006).
90. Y. Kashi, D. G. King, Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* **22**, 253–259 (2006).
91. A. J. Betancourt, K. H. Wei, Y. Huang, Y. C. G. Lee, Causes and consequences of varying transposable element activity: An evolutionary perspective. *Annu. Rev. Genomics Hum. Genet.* **25**, 1–25 (2024), 10.1146/annurev-genom-120822-105708.
92. X. Liao *et al.*, Repetitive DNA sequence detection and its role in the human genome. *Commun Biol.* **6**, 954 (2023).
93. L. Azevedo, C. Serrano, A. Amorim, D. N. Cooper, Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. *Hum. Genomics* **9**, 21 (2015).
94. D. Porubsky *et al.*, Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005.e1926 (2022).
95. R. Villoutreix *et al.*, Inversion breakpoints and the evolution of supergenes. *Mol. Ecol.* **30**, 2738–2755 (2021).
96. A. M. Westram, R. Faria, K. Johannesson, R. Butlin, N. Barton, Inversions and parallel evolution. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **377**, 20210203 (2022).
97. E. Durmaz, E. Kerdaffrec, G. Katsianis, M. Kapun, T. Flatt, "How selection acts on chromosomal inversions" in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd., 2020), pp. 307–315, 10.1002/9780470015902.a0028745.

98. R. Faria, K. Johannesson, R. K. Butlin, A. M. Westram, Evolving inversions. *Trends Ecol. Evol.* **34**, 239–248 (2019).
99. M. Kapun, D. K. Fabian, J. Goudet, T. Flatt, Genomic evidence for adaptive inversion clines in *Drosophila melanogaster*. *Mol. Biol. Evol.* **33**, 1317–1336 (2016).
100. M. Kirkpatrick, How and why chromosome inversions evolve. *PLoS Biol.* **8**, e1000501 (2010).
101. L. Gozashti, O. S. Harringmeyer, H. E. Hoekstra, How repeats rearrange chromosomes in deer mice. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.05.29.596518 (Accessed 1 June 2024).
102. A. A. Hoffmann, L. H. Rieseberg, Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* **39**, 21–42 (2008).
103. M. Recuerda, L. Campagna, How structural variants shape avian phenotypes: Lessons from model systems. *Mol. Ecol.* **33**, e17364 (2024), 10.1111/mec.17364.
104. Y. Wang *et al.*, Transcriptome analysis of comb and testis from Rose-comb Silky chicken (R1/R1) and Beijing Fatty wild type chicken (r/r). *Poult. Sci.* **96**, 1866–1873 (2017).
105. E. R. Funk *et al.*, A supergene underlies linked variation in color and morphology in a Holarctic songbird. *Nat. Commun.* **12**, 6833 (2021).
106. I. Sanchez-Donoso *et al.*, Massive genome inversion drives coexistence of divergent morphs in common quails. *Curr. Biol.* **32**, 462–469.e466 (2022).
107. M. Puig, S. Casillas, S. Villatoro, M. Cáceres, Human inversions and their functional consequences. *Briefings Funct. Genomics* **14**, 369–379 (2015).
108. O. S. Harringmeyer, H. E. Hoekstra, Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nat. Ecol. Evol.* **6**, 1965–1979 (2022), 10.1038/s41559-022-01890-0.
109. G. A. Bravo, C. J. Schmitt, S. V. Edwards, What have we learned from the first 500 avian genomes? *Annu. Rev. Ecol. Evol. Syst.* **52**, 611–639 (2021).
110. M. D. Carling, R. T. Brumfield, Gene sampling strategies for multi-locus population estimates of genetic diversity (θ). *PLoS ONE* **2**, e160 (2007).
111. J. Felsenstein, Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Mol. Biol. Evol.* **23**, 691–700 (2005).
112. J. Ebler *et al.*, Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
113. G. Hickey *et al.*, Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
114. E. Garrison, Z. N. Kronenberg, E. T. Dawson, B. S. Pedersen, P. Prins, A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput. Biol.* **18**, e1009123 (2022).
115. P. Danecek *et al.*, Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
116. Y. H. Liu, C. Luo, S. G. Golding, J. B. Ioffe, X. M. Zhou, Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data. *Nat. Commun.* **15**, 2447 (2024).
117. X. Zheng *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
118. G. Zhang *et al.*, Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
119. S. Purcell *et al.*, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
120. H. Li *et al.*, A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
121. B. Fang, S. V. Edwards. Haemorhous mexicanus (house finch). NCBI SRA. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1101522. Deposited 21 October 2024.
122. B. Fang, house-finch-pangenome. Github. https://github.com/fangbohao/house-finch-pangenome. Deposited 2 October 21 2024.
123. B. Fang, S. V. Edwards, Data from "Fitness consequences of structural variation inferred from a House Finch pangenome." Dryad. https://doi.org/10.5061/dryad.hhmgqnkqb. Deposited 2 October 21 2024.