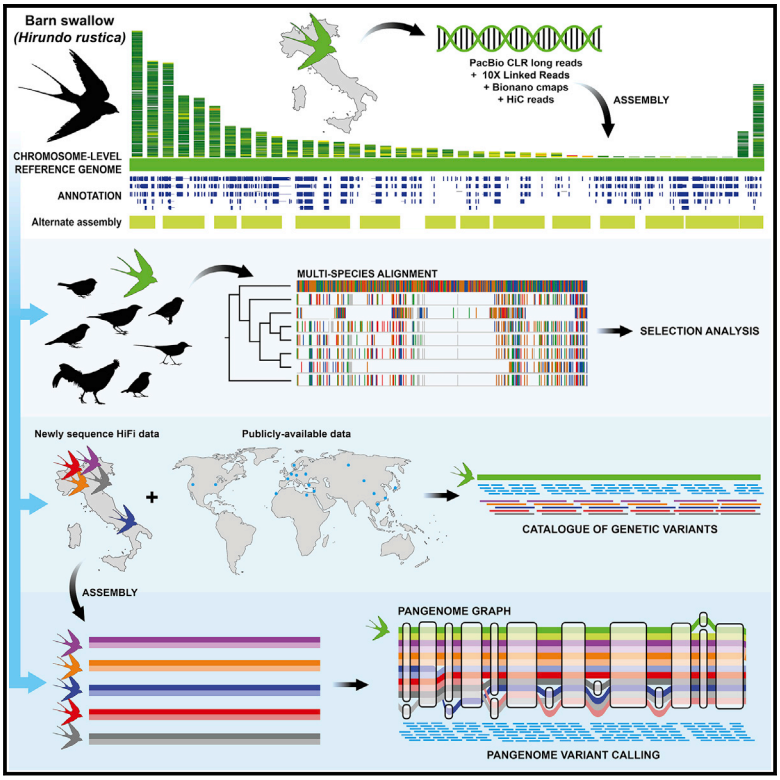
Cell Reports

Resource

A chromosome-level reference genome and pangenome for barn swallow population genomics

Graphical abstract



Highlights

- Generation of a high-quality annotated reference genome and pangenome for barn swallow
- Generation of comprehensive barn swallow genetic variants catalog
- Multispecies alignment and variants catalog detected list of candidate genes
- Pangenome improves read mapping and variant calling

Authors

Simona Secomandi, Guido R. Gallo, Marcella Sozzoni, ..., Erich D. Jarvis, Luca Gianfranceschi, Giulio Formenti

Correspondence

luca.gianfranceschi@unimi.it (L.G.), gformenti@rockefeller.edu (G.F.)

In brief

Secomandi et al. present a chromosome-level genome and pangenome for the barn swallow. They generate a large catalog of worldwide genetic variants and identify genomic regions potentially under selection. They also compare the barn swallow genome with that of other bird species to detect conserved and accelerated genes.

Secomandi et al., 2023, Cell Reports *42*, 111992
January 31, 2023
https://doi.org/10.1016/j.celrep.2023.111992

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

A chromosome-level reference genome and pangenome for barn swallow population genomics

Simona Secomandi,[1,2,17] Guido R. Gallo,[1,17] Marcella Sozzoni,[1] Alessio Iannucci,[3] Elena Galati,[1] Linelle Abueg,[4] Jennifer Balacco,[4] Manuela Caprioli,[5] William Chow,[6] Claudio Ciofi,[3] Joanna Collins,[6] Olivier Fedrigo,[4] Luca Ferretti,[7] Arkarachai Fungtammasan,[8] Bettina Haase,[4] Kerstin Howe,[6] Woori Kwak,[9] Gianluca Lombardo,[7] Patrick Masterson,[10] Graziella Messina,[1] Anders P. Møller,[11] Jacquelyn Mountcastle,[4] Timothy A. Mousseau,[12] Joan Ferrer Obiol,[5] Anna Olivieri,[7] Arang Rhie,[13] Diego Rubolini,[5] Marielle Saclier,[1] Roscoe Stanyon,[3] David Stucki,[14] Françoise Thibaud-Nissen,[10] James Torrance,[6] Antonio Torroni,[7] Kristina Weber,[14] Roberto Ambrosini,[5] Andrea Bonisoli-Alquati,[15] Erich D. Jarvis,[4,16] Luca Gianfranceschi,[1,*] and Giulio Formenti[4,18,*]

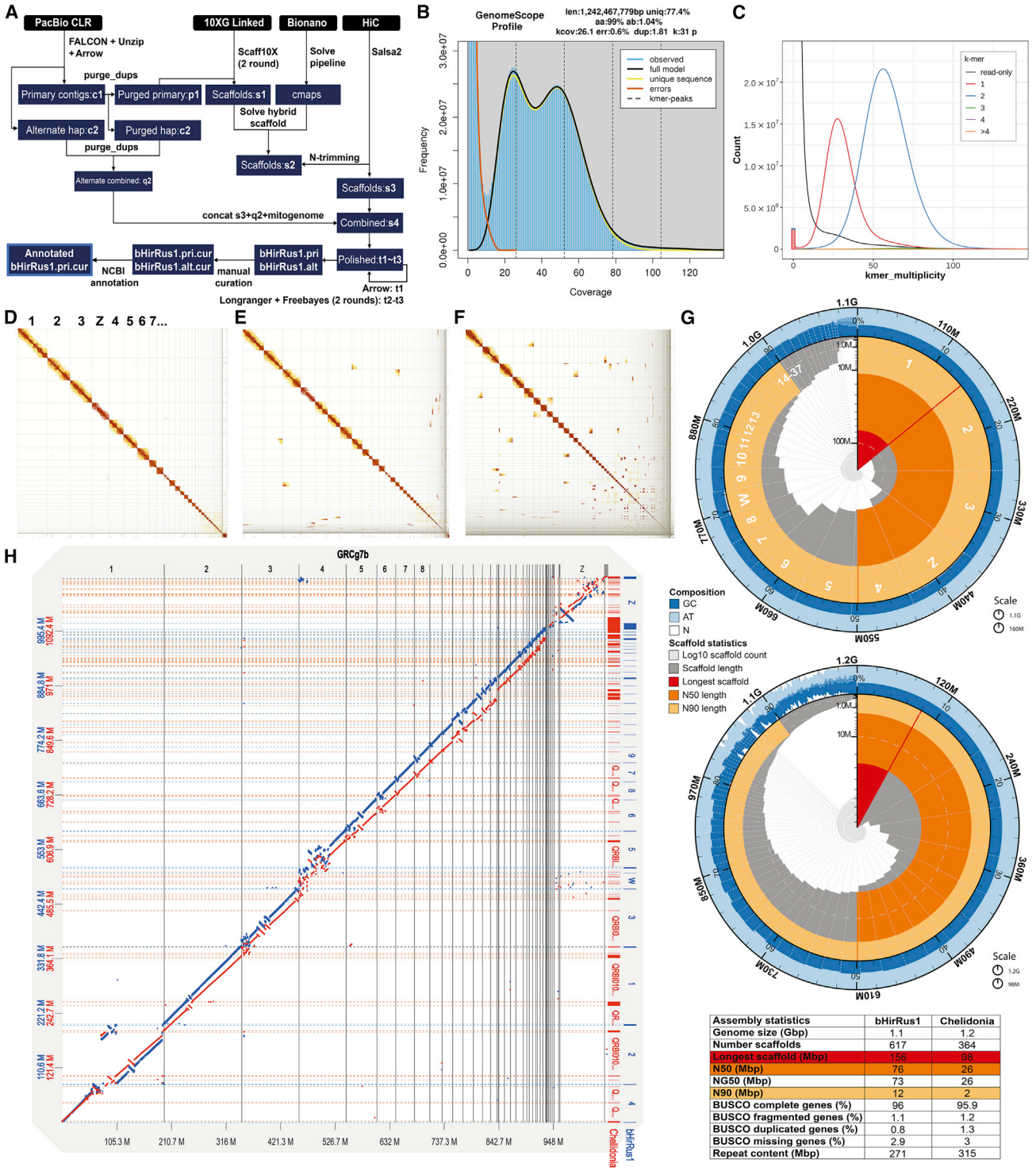
[1]Department of Biosciences, University of Milan, Milan, Italy
[2]Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus
[3]Department of Biology, University of Florence, Sesto Fiorentino (FI), Italy
[4]Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA
[5]Department of Environmental Sciences and Policy, University of Milan, Milan, Italy
[6]Wellcome Sanger Institute, Cambridge, UK
[7]Department of Biology and Biotechnology "L. Spallanzani", University of Pavia, Pavia, Italy
[8]DNAnexus, Inc., Mountain View, CA, USA
[9]Department of Medical and Biological Sciences, The Catholic University of Korea, Bucheon 14662, Korea
[10]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[11]Ecologie Systématique Evolution, Université Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, Orsay Cedex, France
[12]Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA
[13]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
[14]Pacific Biosciences, Menlo Park, CA, USA
[15]Department of Biological Sciences, California State Polytechnic University - Pomona, Pomona, CA, USA
[16]The Howard Hughes Medical Institute, Chevy Chase, MD, USA
[17]These authors contributed equally
[18]Lead contact
*Correspondence: luca.gianfranceschi@unimi.it (L.G.), gformenti@rockefeller.edu (G.F.)
https://doi.org/10.1016/j.celrep.2023.111992

SUMMARY

Insights into the evolution of non-model organisms are limited by the lack of reference genomes of high accuracy, completeness, and contiguity. Here, we present a chromosome-level, karyotype-validated reference genome and pangenome for the barn swallow (*Hirundo rustica*). We complement these resources with a reference-free multialignment of the reference genome with other bird genomes and with the most comprehensive catalog of genetic markers for the barn swallow. We identify potentially conserved and accelerated genes using the multialignment and estimate genome-wide linkage disequilibrium using the catalog. We use the pangenome to infer core and accessory genes and to detect variants using it as a reference. Overall, these resources will foster population genomics studies in the barn swallow, enable detection of candidate genes in comparative genomics studies, and help reduce bias toward a single reference genome.

INTRODUCTION

The barn swallow (*Hirundo rustica*) is an abundant and charismatic migratory passerine bird with six recognized subspecies in Europe, Asia, Africa, and the Americas.[1] Recent reconstructions of its demographic history based on genomic data suggest that its current distribution derives from a relatively recent expansion. The expansion was driven by the spread of human settlements, providing more nesting opportunities[2,3] and leading to the onset of synanthropic habits in this species (i.e., when a species lives in areas occupied and altered by humans).[4,5] Although a large number of studies have focused on barn swallow behavior[6–8] and ecology,[6,8–11] the investigation of phenotype-genotype relationships has been limited by the lack of a highly contiguous, complete, and well-annotated reference genome.[12,13] Two fragmented assemblies for the barn swallow based on short reads were generated in 2016 (*H. r. erythrogaster*)[14] and 2020 (*H. r. rustica*),[15] respectively, while the first reference genome based on long reads was released in 2019 by our research group.[16] The latter is a scaffold-level assembly for



Assembly statistics	bHirRus1	Chelidonia
Genome size (Gbp)	1.1	1.2
Number scaffolds	617	364
Longest scaffold (Mbp)	156	98
N50 (Mbp)	76	26
NG50 (Mbp)	73	26
N90 (Mbp)	12	2
BUSCO complete genes (%)	96	95.9
BUSCO fragmented genes (%)	1.1	1.2
BUSCO duplicated genes (%)	0.8	1.3
BUSCO missing genes (%)	2.9	3
Repeat content (Mbp)	271	315

Figure 1. A *de novo* chromosome-level reference genome for the barn swallow

(A) Flowchart of the VGP assembly pipeline 1.6 (redrawn from Rhie et al.[12]).

(B) Genomescope2.0[21] *k*-mer profile for bHirRus1 generated from trimmed 10x Linked-Reads, used to estimate genome size, repetitiveness, and heterozygosity (top). The x axis represents multiplicity in the read set, while the y axis represents their cumulative frequency.

(C) Merqury[25] spectra-cn plots for bHirRus1. *K*-mer multiplicity in the 10x Linked-Reads (x axis) versus their frequency (y axis). Colored curves discriminate *k*-mer occurrences in the assembly. The bar at the origin of the graph represents *k*-mers found only in the assembly (assembly errors). Two frequency peaks are visible: a haploid peak at ~25× coverage (half average coverage, red), representing *k*-mers found once in the assembly (haplotype specific), and a diploid peak at ~50×

(legend continued on next page)

the *H. r. rustica* (the Eurasian subspecies) generated by combining PacBio long-read sequencing[17] and Bionano Direct Label and Stain (DLS) optical mapping.[18] Here we present the first chromosome-level reference genome for the same individual[16] generated using the Vertebrate Genomes Project (VGP) assembly pipeline.[12] With this reference genome we identified conserved and accelerated regions in the barn swallow genome and generated a catalog of genetic markers using all publicly available data to accurately estimate linkage disequilibrium (LD). Genome-wide analyses led to a list of candidate genes potentially under selection in this species. Recently, algorithmic advances have led to the concept of pangenome reference graphs, which promise to improve variant calling, a pivotal requirement for phenotype-genotype association studies.[19,20] Therefore, we also present the first pangenome graph for the barn swallow. We tested its use for read mapping and variant calling, highlighting the potential of pangenome graphs for population genomics.

RESULTS AND DISCUSSION

A new reference genome for the barn swallow

Using the VGP genome assembly pipeline v.1.6[12] (Figure 1A), we generated the first chromosome-level reference genome ("bHirRus1" hereafter) and an alternative-haplotype assembly for the barn swallow. Contigs were generated using PacBio CLR long reads and scaffolded with 10x Linked-Reads, Bionano optical maps, and Hi-C reads. We also generated a draft mitochondrial genome for the species (Figure S1; Data S1). We sequenced a female (the heterogametic sex) to obtain both sex chromosomes. After manual curation (Figure 1D; and see Figure 1E and Data S1), the primary assembly is 1.11 gigabase pairs (Gbp) long, close to Genomescope2.0[21] predictions (Figure 1B; Tables S1A and S1B; Data S1). The assembly has a scaffold NG50 of 73 megabase pairs (Mbp), a per-base consensus accuracy (QV) of 43.7 (~0.42 base errors/10 kilobase pairs [kbp]) and a *k*-mer completeness of 83.3% with a duplication content of 0.49% (Figures 1C and 1G; Tables S1B and S1C; Data S1). Functional gene completeness, measured with BUSCO,[22] is 96% (Figure 1G; Table S1D). We assigned 98.2% of the assembled sequence to 39 autosomes and to the Z and W sex chromosomes (Figure 1G; Table S2), which are usually challenging to assemble due to their highly repetitive nature.[23] The assembly exceeds the VGP standard metrics (6.7.Q40.C90).[12] The chromosome reconstruction (2n = 80) matches our cytogenetic analysis (Figure 2A; Data S1), in line with the current literature on pachytene karyotypes for the barn swallow.[24] We defined chromosomes 1–6 and Z as macrochromosomes, 7–13 and W as intermediate chromosomes, and 14–39 as microchromosomes (Data S1). The size of the assembled chromosome sequences tightly correlates with the physical size of the chromosomes, estimated from karyotype images (Spearman's $\rho = 0.99$, $n = 40$, $p < 2.2 \times 10^{-16}$; Figure 2B; Table S3). As expected,[12] PacBio long reads show haploid coverage for Z and W (Figure 2C, track A). The total repeat content of bHirRus1 is 271 Mbp (22.9%; Figure 2C, track B; Table S2), in line with Genomescope2.0[21] predictions (Figure 1B; Table S1A), while the GC content is 42.5% (Figure 2C, track C; Table S2).

Functional annotation

Using newly generated and already available transcriptomic data (Table S4A), we used the NCBI Eukaryotic genome annotation pipeline[12,27] to identify 18,578 genes and pseudogenes, 15,516 of which are protein coding. Among these, 15,130 (97.5%) align to UniProtKB/Swiss-Prot-curated proteins, covering ≥50% of the query sequence, while 10,797 (69.6%) coding sequences align for ≥95%. In line with other birds,[28] ~52% of the total bp is annotated as genes, of which ~90% are annotated as introns and ~5% as coding sequences (CDSs; Table S4B).

Chromosome size and genomic content

Differences in GC, CpG islands, gene and repeat content between birds' chromosome types are likely the product of the evolutionary process that led to stable chromosome classification in birds.[29] Similar to the zebra finch (*Taeniopygia guttata*) genome,[30] bHirRus1 chromosome size negatively correlates with GC content (Spearman's $\rho = -0.972$, $n = 38$, $p < 2.2 \times 10^{-16}$); CpG island density (Spearman's $\rho = -0.925$, $n = 38$, $p < 2.2 \times 10^{-16}$); gene density (Spearman's $\rho = -0.364$, $n = 38$, $p < 2.5 \times 10^{-2}$); and repeat density (Spearman's $\rho = -0.51$, $n = 38$, $p = 1.2 \times 10^{-3}$; Figure 2C, tracks B–E; Table S2). Indeed, microchromosomes are GC rich

(average coverage, blue) representing *k*-mers found twice in the assembly (shared between haplotypes). No *k*-mers resulting from artificial duplications (green, purple, yellow) are visible (duplication content 0.49%; Table S1).

(D) Hi-C interaction heatmaps for the curated bHirRus1 assembly. The linear sequence of the reference genome assembly is represented on both axes, and the diagonal shows 3D proximity of interacting pairs. The strength of the interaction is given by color intensity. A scaffold is considered a full chromosome when the number of interchromosomal interactions is negligible. No off-diagonal interactions are visible. Scaffolds are labeled by their chromosome number.

(E) Hi-C interaction heatmaps for bHirRus1 assembly before curation. A number of off-diagonal interactions are still visible, which can either result from missing links between scaffolds of the same chromosome or from misassembly.

(F) Hi-C interaction heatmaps for Chelidonia assembly. The assembly is still substantially fragmented, with several off-diagonal Hi-C interactions.

(G) Snail plots and assembly summary statistics. The main plot is divided into 1,000 size-ordered bins around the circumference. Scaffold length distribution is shown in dark gray with the plot radius scaled to the longest scaffold (red). Orange and pale orange arcs show scaffold N50 and N90, respectively. The pale gray spiral shows the cumulative scaffold count on a log scale, with white scale lines showing successive orders of magnitude. The blue and pale blue areas around the plot show the GC, AT, and N content in the same bins as the inner plot. Top plot: bHirRus1 snail plot. Bottom plot: Chelidonia snail plot. The table summarizes the assembly summary statistics and BUSCO[26] results (vertebrata_odb10) of Chelidonia and bHirRus1.

(H) Dotplot alignment of bHirRus1 (blue) and Chelidonia (red) with the VGP chicken assembly GRCg7b. Chromosome numbers and coordinates are reported for GRCg7b (x axis), Chelidonia (y axis, red), and bHirRus1 (y axis, blue). Black vertical lines, red horizontal lines, and blue dashed horizontal lines define chromosome and scaffold boundaries in the chicken assembly, in Chelidonia, and in bHirRus1, respectively.

See also Figure S10 and Table S1.

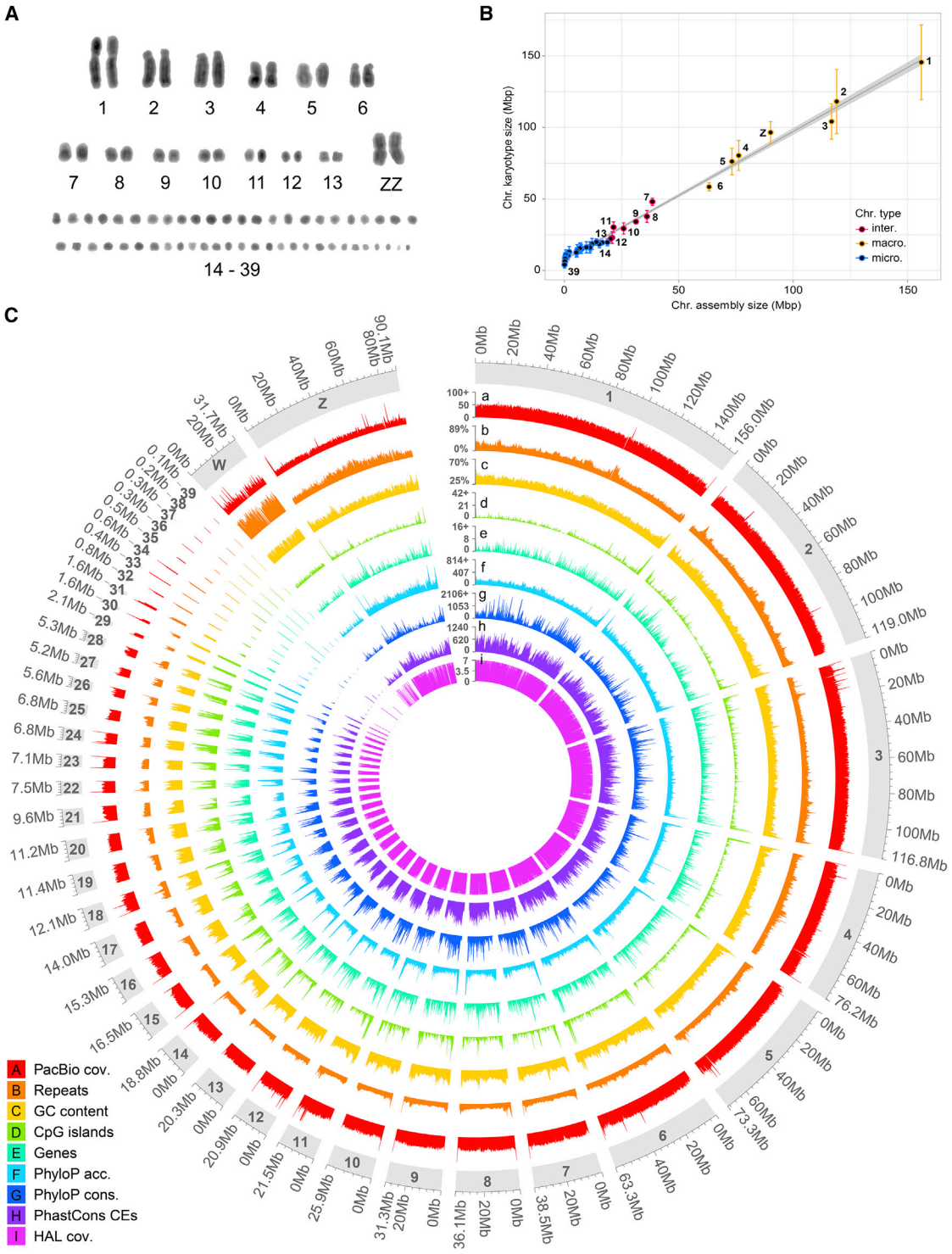


Figure 2. Karyotype reconstruction and reference genome chromosome characteristics

(A) 4′,6-diamidino-2-phenylindole (DAPI)-stained karyotype of a male *H. r. rustica* individual (inverted colors).

(B) Correlation between assembled chromosome length (x) and the estimated chromosome length from karyotype images (y). The W sex chromosome is absent due to the sex of the karyotyped sample.

(C) Circular representation of bHirRus1 chromosomes. All data are plotted using 200 kbp windows, and the highest values were capped at the 99% percentile value for visualization whenever necessary (marked with +). PacBio long-read coverage (a); percentage of repeat density (b); percentage of GC (c); CpG island density (d); gene density (e); phyloP accelerated site density (f); phyloP conserved site density (g); phastCons conserved element (CE) density (h); and coverage of bHirRus1 in the Cactus HAL alignment (i).

See also Figures S2 and S3 and Tables S2, S3, S5, and S6.

(Mann-Whitney U test, W = 0, p = 2.8 × 10^{-7}); CpG rich (Mann-Whitney U test, W = 3, p = 4.5 × 10^{-7}); gene rich (Mann-Whitney U test, W = 94, p = 2 × 10^{-2}); and repeat rich (Mann-Whitney U test, W = 103, p = 3.9 × 10^{-2}).

Comparison between bHirRus1 and previous assemblies

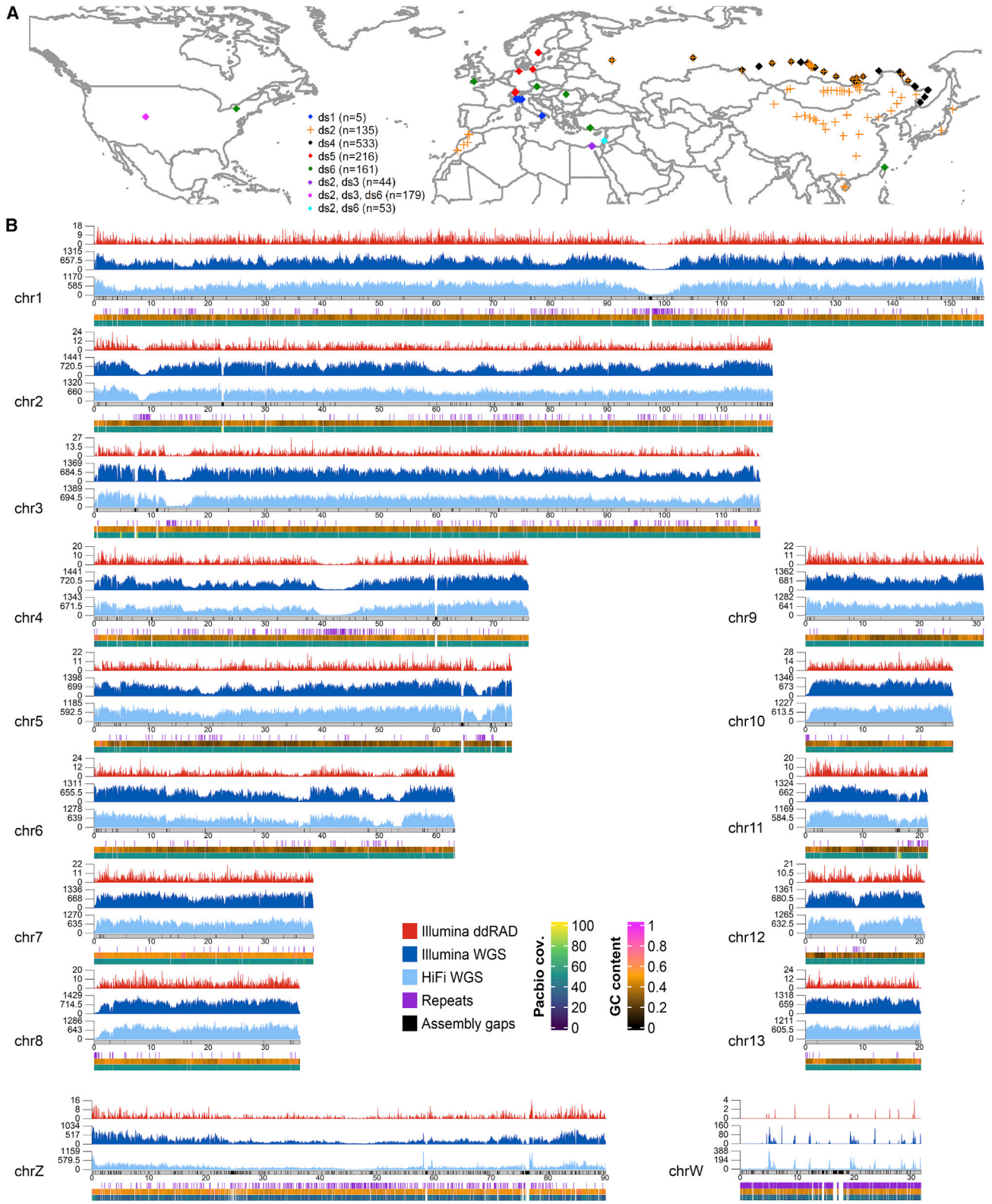
Two previous barn swallow genome assemblies, based on short reads, were released in 2016 and 2020. They showed a contig N50 of 39 kbp[14] and a scaffold NG50 of 676 kbp,[15] respectively, considerably lower than bHirRus1 (contig N50: 2.8 Mbp; scaffold NG50: 73 Mbp; Table S1B). With respect to the 2020 assembly, bHirRus1 showed a higher quality and completeness (BUSCO score: 96% vs. 53.8%, QV: 43.7% vs. 24.3%, *k*-mer completeness: 83.3% vs. 40.3%; Tables S1C and S1D). With respect to the 2019 long-read-based assembly[16] (here after "Chelidonia"), the VGP assembly pipeline and our subsequent manual curation increased the assembly contiguity to the chromosome level (scaffold NG50: 26 vs. 73 Mbp; Figure 1G; Table S1B; see Data S1 for the expanded comparison). The higher contiguity of bHirRus1 is also confirmed by the Hi-C contact heatmap (Figures 1D vs. 1F), a data type previously unavailable,[16] and by the alignment with the chicken genome GRCg7b (Figure 1H). Assembly QV also considerably increased in bHirRus1 (43.7 vs. 34; Table S1C). The repeat content decreased from 315 to 271 Mb (Figure 1G). BUSCO completeness slightly increased in bHirRus1 (96% vs. 95.9%), with less duplicated (0.8% vs. 1.3%) and marginally less fragmented (1.1% vs. 1.2%; Figure 1G; Table S1D) BUSCO genes. Overall, our results confirm the need for long reads and physical information in genome assembly to increase contiguity and completeness.[12,31]

Reference-free, whole-genome multiple species alignment and selection analysis

To identify regions under positive selection (i.e., evolving at a higher rate than under neutral evolution) and under negative selection (i.e., evolving at a lower rate), we generated a reference-free, whole-genome multiple alignment using Cactus.[32] The alignment included bHirRus1, six publicly available chromosome-level Passeriformes genomes, and the chicken GRCg7b genome (Figure S3A; Table S5A). The coverage of the alignments with bHirRus1 (mean alignability: 76%; Table S5A) was uniform, with the exception of chromosome W and the smallest microchromosomes (Figure 2C, track I; Table S5B). Using a 4-fold-degenerate sites neutral model and the Cactus alignment in phyloP,[33] we found that 0.96% of bHirRus1 bases are accelerated and 2.71% are conserved after false discovery rate (FDR) correction[34] (Figures 1C, tracks F and G, S3B, S3C, S3E, and S3F; Table S6A). Using phastCons,[35] we identified ~3 million conserved elements (CEs) covering 12.3% of the barn swallow genome (133 Mbp; Figure 2C, track H; Table S6A). Among the accelerated and conserved bases detected by phyloP, about 52% and 63%, respectively, fall within genes, while only ~0.9% and ~17% overlapped with CDSs, in line with previous studies[36,37] (Figure S3D; Table S6B). PhastCons CEs showed similar overlaps (genes: ~61%, CDSs: ~14%; Figure S3D; Table S6B). PhyloP conserved sites positively correlated with phastCons CEs (Spearman's ρ = 0.83, n = 108,010, p < 2.2 × 10^{-16}). Based on our results, phyloP sites can be considered a higher confidence subset within the larger phastCons set (see Figure S4 for an example), and we therefore based our subsequent analyses on phyloP results. Conserved site density was weakly positively correlated with chromosome sizes (Spearman's ρ = 0.35, n = 38, p < 3.4 × 10^{-2}) without significant differences between chromosome types (Wilcoxon test, W = 244, p = 0.189). Conversely, accelerated site density was strongly negatively correlated with chromosome size (Spearman's ρ = −0.80, n = 38, p < 9.5 × 10^{-8}), with microchromosomes richer in accelerated sites than other chromosome types (Wilcoxon test, W = 50, p = 4.6 × 10^{-5}), as already observed in other birds.[38] Gene Ontology (GO) analysis on the top 5% of genes with highest overlap with phyloP accelerated sites (Table S7) did not disclose any enriched GO term (Table S8; Data S1). As expected, we detected an enrichment of conserved bases in CDSs compared with the non-coding regions of genes[15] (χ^2 = 2.03 × 10^7, df = 1, p < 2 × 10^{-16}). The GO analysis on the top 5% of genes with the largest number of phyloP conserved sites within the CDS (Table S9) revealed an enrichment for genes involved in DNA binding, transcriptional regulation, and nervous system development (Table S10). The top 20 conserved genes are largely involved in neural development and differentiation (Table S9; Data S1). Among the top six, we found genes involved in stress-related pathways (*camk2n2*, *inhbb*, *sumo2*, *nfia*, *sox2*, *cnot*; see Data S1 for more details on gene functions and an additional analysis regarding *camk2n2* potential involvement in the onset of synanthropic behaviors). The top candidate, *camk2n2*, located on chromosome 10, has the same base composition in the CDS in all species, with the exception of the chicken, which has few single-nucleotide polymorphisms (SNPs; 3 SNPs in the first CDS, 1 in the second CDS; Figure S4). The variability increases when considering non-coding regions (Figure S4). The conserved genes detected by phyloP analysis deserve further study as candidate genes, likely providing insights into the pathways and functions potentially under selection.

Marker catalog and genome-wide density

To obtain a comprehensive catalog of SNPs (Data S1), we generated high-coverage HiFi data (ds1, ~20× coverage, n = 5) for five *H. r. rustica* individuals (Table S11A) and aligned them using bHirRus1 as reference. We complemented this information with all the publicly available genomic data for the species (Figure 3A; Table S12), including two Illumina whole-genome sequencing (WGS) datasets[2,39] (ds2 and ds3.1, ~6.8×, n = 159) and four ddRAD datasets[2,14,40,41] (ds3.2 through ds6, ~0.07×; n = 1,162). Despite the fewer individuals in HiFi WGS, the average SNP density and distribution (Figures 3B and S5, light blue track; 142.37 SNPs/10 kbp; Table S13) was comparable to the one computed for Illumina WGS (Figures 3B and S5, dark blue track; 160.34 SNPs/10 kbp; Table S13). Since read accuracy of the two systems is very similar (99.9%), we hypothesized that the higher number of variants per sample was due to the higher read mappability of HiFi reads spanning complex genome regions. We also performed a coverage titration experiment (Data S1) and found that SNP distribution was still uniform across chromosomes even when HiFi WGS was downsampled



(legend on next page)

to 5× (96.33 SNPs/10 kbp; Figure S6; Table S13), supporting our hypothesis. Chromosome W showed the lowest SNP density among all chromosomes (HiFi WGS: 3.16 SNPs/10 kbp; HiFi WGS: 5× 1.01 SNPs/10 kbp; Illumina WGS: 1.38 SNPs/10 kbp), in line with the facts that it is present as a single copy only in females and that it has the highest content of heterochromatin and repeat elements, hindering variant calling.[42] In contrast, we identified a higher number of SNP markers on chromosome Z (HiFi WGS: 31.8 SNPs/10 kbp; HiFi WGS: 5× 2.34 SNPs/10 kbp; Illumina WGS: 53.3 SNPs/10 kbp). As expected, ddRAD exhibited very localized peaks of SNPs (0.8 SNPs/10 kbp; Figures 3B and S5, red track). Particularly, ddRAD identified an extremely low number of SNPs on chromosome Z (0.27 SNPs/10 kbp) and no SNPs on microchromosome 33 (Figure S5). As observed in other bird species,[43,44] we detected a positive correlation between chromosome GC content and SNP density in all datasets (Data S1).

Genome-wide LD

A comprehensive set of genetic markers accurately mapped on a high-quality assembly represents a suitable resource for several population genomics analyses. The power and precision of association mapping and quantitative trait loci (QTLs) detection depend on LD,[45] and assessing its decay is pivotal to the success of genome-wide association studies (GWAS).[46,47] To this end, we assessed genome-wide LD decay using the SNPs in our catalog derived from Illumina WGS (ds2 and ds3.1). We found that genome-wide average r^2 varied between *H. rustica* subspecies (Figure 4A; Table S14). As expected,[48] absolute r^2 decreased with increasing sample size and marker distance (Figure 4A; Table S14). Overall, our results indicate that the genetic association between loci in the barn swallow is extremely low and decreases rapidly within the first 10 kbp, as expected in large panmictic populations.[49] Indeed, no evidence of population structure has been observed in the European subspecies (*H. r. rustica*), potentially due to extensive gene flow between breeding populations.[40] Average r^2 at increasing distance varied also across chromosome types, confirming that avian microchromosomes are characterized by higher rates of meiotic recombination, resulting in lower LD, than macrochromosomes (Figure 4B; Table S15).[29,50,51] Additionally, a chromosome scan for high-LD regions, allowed by dense SNP catalogs such as the one presented here, led to the identification of genes putatively under selection (please refer to Data S1 for a detailed analysis of the top candidate genes, including *bdnf* and *lgr4*).

Toward a pangenome for the barn swallow

Despite the high resolution achieved with chromosome-level assemblies, population genomic studies based on traditional linear reference genomes face limitations when aiming to describe complete variation among individuals.[19,20] To reduce bias toward a single reference genome in future studies, we assembled our newly generated high coverage HiFi data (ds1) with Hifiasm[52] and used both primary and alternate haplotypes (Table S11C), together with bHirRus1 primary and alternate assemblies, to generate the first pangenome graph[53,54] for the species (Figure 5). All the HiFi individuals, considering both haplotypes, shared 92.6% of bHirRus1 genes (core genes; Figures 5A and 5B; Table S16). 1.29% (234) were not found in the HiFi assemblies (putative bHirRus1 accessory genes; Figure 5B; Tables S16 and S17). Of those genes, 79 were found in the HiFi raw reads of at least one individual for >80% of their sequence with >99% identity, lowering the number of the putative bHirRus1 accessory genes from 234 to 155 (0.85%; Figure 5C; Table S17). 106 out of the 155 genes absent from both HiFi raw reads and HiFi-based assemblies belong to unlocalized or unplaced scaffolds in bHirRus1 (Table S17), suggesting that these genes may have also been hard to sequence and assemble in the reference. The 155 missing genes are enriched in GC content compared with the rest of bHirRus1 genes (Mann-Whitney U test, $W = 709{,}383$, $p < 2.2 \times 10^{-16}$; Figure 5D; Table S17). By measuring the percentage of 128 bp windows with >50% dinucleotide composition, we also found a significant enrichment in GC (2.6% vs. 0.9%; $\chi^2 = 601.8$, df = 1, $p < 0.0001$) and GA dinucleotides (2.3% vs. 1%; $\chi^2 = 315.7$, df = 1, $p < 0.0001$) and depletion in AT dinucleotides (0.54% vs. 1.5%; $\chi^2 = 115.7$, df = 1, $p < 0.0001$; Figure 5E; Table S18). GA dinucleotide enrichment has been described as particularly challenging for several polymerase enzymes, including the one used in PacBio sequencing.[55–57] This suggests that further validation and additional data are warranted to accurately characterize the core and accessory genome of the barn swallow.

We then focused on the top conserved candidate gene *camk2n2* region in the pangenome. Similar to what we had observed between species (Figure S4), we found high conservation of the two CDSs among the five barn swallow individuals (Figure 5F; see Figure S7A for a zoom on the CDS). We detected 60 SNPs in non-coding regions (Figure 5F), confirming a higher variability than in CDSs (1 SNP) within the same species, in line with what we observed between species (Figure S4). To confirm these SNPs, we examined the raw calls obtained from HiFi reads (ds1) mapped against our linear reference genome. The calls

Figure 3. Sampling locations and SNP density per chromosome

(A) Sampling locations of all individuals used to generate the SNP catalog. Purple, fuchsia, and light blue colors indicate sampling locations in common between datasets indicated in the legend. Sampling locations from ds2 are plotted with a different shape (cross) to distinguish them from black points (ds4), as some sampling locations partially overlap on the map. Data of populations of ds2 through ds6 are from publicly available genomic data.

(B) Only macrochromosomes and intermediate chromosomes are shown. Microchromosomes are shown in Figure S5. SNP density was computed over 40 kbp windows. Numbers on the y axis of each density track indicate the maximum and average values of SNP density for each track. Genomic data types are color coded. Light blue: HiFi WGS data (ds1). Dark blue: Illumina WGS data from ds2 and ds3.1. Red: Illumina ddRAD data from ds3.2 through ds6.8. All available samples from the same sequencing technology were considered together. Additional tracks in the bottom panel show repetitive regions of the genome (violet bars; only regions larger than 3 kbp are plotted), GC content, and PacBio reads coverage. Gray ideograms represent chromosomes in scale, with assembly gaps highlighted as black bars.

See also Figures S5 and S6 and Tables S11, S12, and S13.

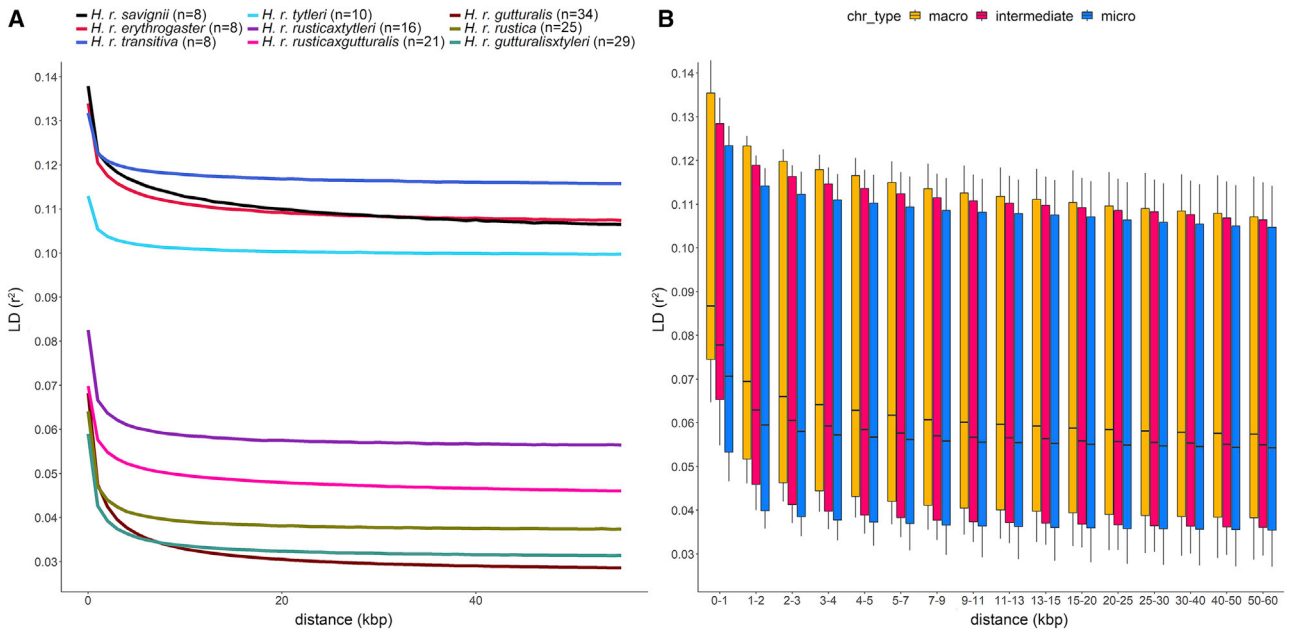


Figure 4. Linkage disequilibrium decay in the barn swallow genome

(A) Average r^2 values plotted against physical distance (kbp) for the different populations belonging to ds2 and ds3.1 (Illumina WGS data).

(B) Average r^2 values in macrochromosomes, intermediate chromosomes, and microchromosomes according to pairwise distance (kbp) between SNPs. LD median estimates were obtained averaging values from all Illumina WGS data populations (ds2 and ds3.1).

See also Figure S9 and Tables S14 and S15.

included 53 out of the 60 SNPs detected with the pangenome (Table S19). The missing SNPs were found in the alternate bHirRus1 assembly (Figure 5F), which is present in the pangenome but not considered in single-haplotype reference genome variant calling.[58] To validate variant identification using the pangenome as reference, we mapped the Illumina WGS ds3.1 and called the variants in the *camk2n2* region using vg,[59] comparing them with the variants recovered using bHirRus1 alone. In fact, 8 SNPs were identified from the single reference genome analysis, while the pangenome allowed the recovery of 54 SNPs within the considered region (Table S20). Manual removal of low-confidence variants (STAR Methods) reduced the number of reliable SNPs to 20, comprising all the eight SNPs identified with bHirRus1 (Table S20). A closer inspection of the alignment to the linear genome revealed that 11 of the remaining 12 pangenome variants had support from the reads but were not retained when using Freebayes default parameters. One variant was not supported by any observation from reads aligned to bHirRus1, suggesting that its identification was due to the higher mappability of the reads to the pangenome (Figure S7B; Table S20).

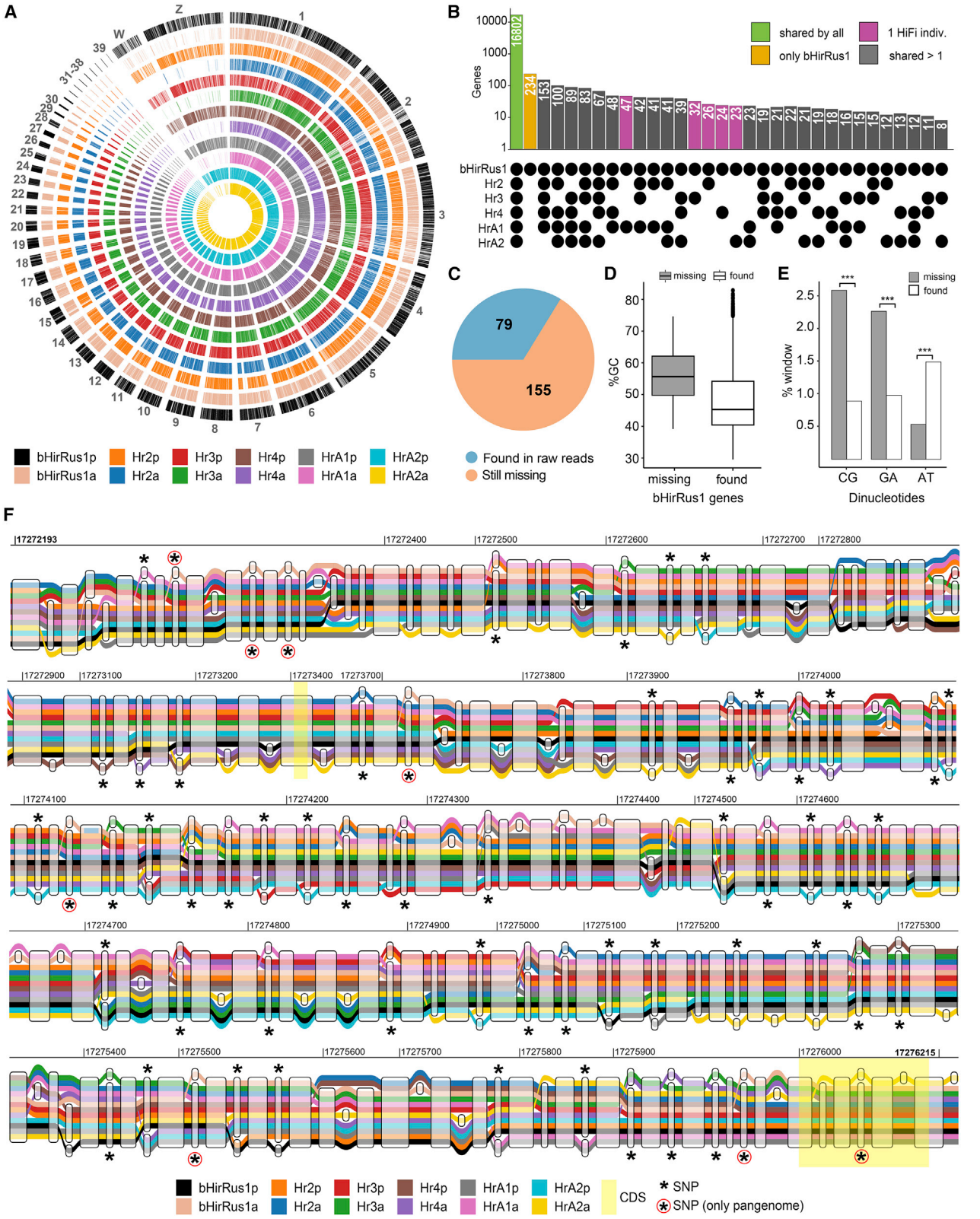
Conclusion

We presented the highest-quality reference genome for the barn swallow, a genome-wide catalog of genetic variants compiled using all publicly available data, and the first pangenome reference graph for the species. A reference genome of such quality allowed us to conduct a wide array of comparative and population genomics analyses, including an accurate estimate of LD patterns in different barn swallow populations, leading to the detection of genomic regions harboring genes potentially implicated in stress response that might have played a role in the evolution of synanthropy[60–64] and song learning.[65] Our pangenome graph constructed from multiple haplotypes allowed us to infer a set of core and accessory genes and also to place variants in the correct haplotype without additional phasing. The use of pangenome graphs promises to improve mappability of resequencing data, avoiding reference bias and ultimately increasing precision and recall rates in population genomic analyses. Our preliminary analyses support this idea, although caution should be used in the interpretation of the results as these new implemented methods still need to be thoroughly validated. Overall, the resources presented here will be instrumental to plan and inform future studies on the barn swallow and other species, including phylogenetic, demographic, and phenotype-genotype association studies.

Limitations of the study

Cactus alignment and selection analysis

With the reference-free alignment we generated using Cactus,[32] we detected conserved and accelerated genes in the barn swallow genome. We are aware that increasing the number of species involved in the alignment would improve the statistical significance of our results.[15] Indeed, due to the low number of aligned species and the low total branch length between them,[15] the basewise selection analysis with phyloP[33,35] failed to detect significant calls after a FDR[34] correction with 0.05 as significance level. We therefore increased the statistical power of the constraint analysis by running the analysis on 10 bp windows. Moreover, we focused on conserved genes and, in particular, on the top candidate *camk2n2*, which may be an interesting gene for the onset of domestic and synanthropic behavior.



(legend on next page)

However, our alignment included species that are all domesticated or somewhat related to human environments, which made it difficult to discern whether the gene is related to domestication and synanthropy or is conserved among all species. Therefore, we only used the gene as an example for the visualization figures (Figures 5F, S4, and S7). Another potential limitation in this analysis is that we could not take into account the heterogeneity in the recombination landscape in birds.[43,44] In the absence of information on the recombination landscape for all the species in the multiple alignment, the current methods cannot account for it, and we therefore avoided speculation about the role of the genes under selection.

Pangenome

The pangenome presented in this publication is the first example in the barn swallow, and it was constructed to show the potential and benefits of using a reference-free genome, compared with a linear reference genome, to call genetic variants. However, we are aware that the relatively small number of individuals used to construct the pangenome, and their inadequate representation of the worldwide variability in the species, may be limitations to its wider use. Nonetheless, we believe that the possibility of integrating the pangenome with new sequence data will facilitate its use and spread, ultimately overcoming the severe limitations of species-specific comparisons associated with a single reference-based approach.

LD scans

With our newly generated chromosome-level reference genome, we investigated the LD decay pattern in different barn swallow populations (Figure 4A) using all WGS data publicly available. The limited sample size (ranging from 8 to 34 per population) should be taken into account when interpreting these results.

We also performed chromosome scans to detect genomic regions with high LD to identify genes putatively under selection. One of the most compelling regions we identified harbors *bdnf*, a very interesting candidate to be considered for future studies (Figure S9; Data S1). We identified a high homozygosity in the genomic region in some of the populations analyzed (Data S1). A potential limitation of our approach might be that we could not take into account the different recombination rate patterns along the barn swallow genome,[66] which play a relevant role in determining homozygosity. Therefore, we cannot exclude that the low diversity observed within this chromosome region could result from low rates of recombination within this genomic region rather than selective pressure only.[67] An alternative possibility is that in the specific case of the Egyptian barn swallow population, where there is evidence of a past bottleneck event,[2] genetic drift might have also played a role in determining this high-LD region. However, we confirmed the presence of a potential selection signature within this genomic region by computing the integrated haplotype homozygosity score (Data S1). Yet, we are aware that these results may not be definitive because of the limited sample size and the partial phasing of genetic variants achievable with short-reads.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Sampling for sequencing
  - Karyotype reconstruction
- METHOD DETAILS
  - DNA extraction
  - Library preparation and sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Mitogenome assembly
  - Reference genome assembly
  - Manual curation
  - Annotation
  - Chromosome size estimations from karyotype images
  - Chromosome classification assignment
  - Assembly evaluation and comparison with other barn swallow assemblies
  - Selection analysis on multiple whole-genome alignments
  - SNP catalog generation

Figure 5. The first pangenome for the barn swallow

(A) Circos plot showing the annotated genes of bHirRus1p (primary assembly) and orthologs found in bHirRus1a (alternate assembly) and the HiFi-based haplotypes.

(B) Histogram reporting presence or absence of bHirRus1 genes in the other individuals of the pangenome (primary and alternate assemblies combined). Green: genes shared by all individuals. Yellow: genes exclusive to bHirRus1. Fuchsia: genes shared between bHirRus1 and another individual. Gray: genes shared between bHirRus1 and 2 or more individuals.

(C) Pie chart reporting the 234 genes exclusive of bHirRus1, i.e., missing from all the other genome assemblies in the pangenome. 79 genes were identified in the HiFi raw reads (light blue), while 155 genes could not be found in either HiFi-based assemblies or HiFi raw reads.

(D) Boxplot representing the GC content among the 155 missing genes from both HiFi assemblies and raw reads (gray) vs. all other bHirRus1 genes (white, found in at least 1 HiFi individual).

(E) Barplot reporting the percentage of 128 bp windows with >50% dinucleotide content in the 155 genes (gray) vs. all other genes (white). The Chi-square analyses were associated with a p value < 0.0001.

(F) Extract of the entire *camk2n2* sequence obtained from the pangenome graph (chromosome 10, 17,272,192–17,276,215 bp). The colored tubes represent the assembled haplotypes included in the pangenome. bHirRus1 Chr10 ("bHirRus1p," black) is shown together with the alternate assembly "bHirRus1a," the five HiFi-based primary assemblies (Hr2p, Hr3p, Hr4p, HrA1p, HrA2p), and their alternate assemblies (Hr2a, Hr3a, Hr4a, HrA1a, HrA2a). CDSs are highlighted with transparent yellow boxes. SNPs are marked with black asterisks. SNPs found with the pangenome, but not detected with the standard variant calling approach, are circled in red.

See also Figures S4 and S7 and Tables S11, S16, S17, S18, and S19.

- Linkage disequilibrium and haplotype statistics analysis
- HiFi reads processing for SNP catalog, titration, and phasing experiment
- Pangenomics
- Comparison between variants embedded in the pangenome and variants called with deepvariant
- Pangenome variant calling
- Graphical representations

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.111992.

ACKNOWLEDGMENTS

This work would have not been possible without the dedication of the late Prof. Nicola Saino. We thank the INDACO Platform team (a project of High Performance Computing at the University of Milan, https://www.indaco.unimi.it/), in particular Dr. Alessio Alessi, as well as Prof. Aureliano Bombarely for providing computational resources and technical assistance. We thank Prof. Guido Grilli (Department of Veterinary, University of Milan, Milan, Italy) and Dr. Alessandra Costanzo for their help in obtaining barn swallow samples. We received support from the Italian Ministry of Education, University and Research (MIUR) for the project PRIN2017 2017CWHLHY (L.G. and A.T.); Dipartimenti di Eccellenza Program (2018–2022) - Department of Biology and Biotechnology "L. Spallanzani'' University of Pavia (to A.O., L.F., and A.T.); the CSU Program for Education & Research in Biotechnology (CSUPERB) (to A.B.-A.); Howard Hughes Medical Institute (to E.D.J.); and the Samuel Freeman Charitable Trust (to T.A.M. and A.P.M.). The work of F.T.-N. and P.M. was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.

AUTHOR CONTRIBUTIONS

S.S., G.R.G., A.I., E.G., J.B., M.C., J.M., M. Saclier, R.S., and G.F. performed the wet-lab experiments. S.S., G.R.G., A.T., A.B.-A., L.G., and G.F. planned the experiments. S.S., G.R.G., M. Sozzoni, A.I., J.F.-O., R.S., P.M., K.W., A.B.-A., L.G., and G.F. analyzed the data. S.S., G.R.G., M. Sozzoni, A.B.-A., L.G., and G.F. drafted the manuscript. C.C., A.P.M., T.A.M., A.T., A.B.-A., E.D.J., and L.G. provided computational resources or funding. S.S., W.C., J.C., K.H., and J.T. performed manual curation. S.S., P.M., and F.T.-N. performed assembly annotation. J.B., O.F., B.H., and J.M. generated the raw sequencing data. S.S. generated the genome assembly with support from A.F. and A.R. S.S., A.I., M.C., D.R., R.A., and G.F. contributed to sampling. S.S., L.A., W.K., E.D.J., and G.F. handled data submission. L.F., G.L., A.O., J.F.-O., D.R., A.T., R.A., A.B.-A., and E.D.J. contributed to the general discussion. All authors reviewed the final manuscript and approved it.

DECLARATION OF INTERESTS

D.S. and K.W. are full-time employees at Pacific Biosciences, a company commercializing single-molecule sequencing technologies.

Received: March 9, 2022
Revised: July 20, 2022
Accepted: January 4, 2023

REFERENCES

1. Spina, F. (1998). The EURING swallow project: a large-scale approach to the study and conservation of a long-distance migrant. Migrating birds know no boundaries. Proc. Int. Symp. Isr. Torgos *28*, 151–162.
2. Smith, C.C.R., Flaxman, S.M., Scordato, E.S.C., Kane, N.C., Hund, A.K., Sheta, B.M., et al. (2018). Demographic inference in barn swallows using whole-genome data shows signal for bottleneck and subspecies differentiation during the Holocene. Mol. Ecol. *27*, 4200–4212. https://doi.org/10.1111/mec.14854.
3. Lombardo, G., Rambaldi Migliore, N., Colombo, G., Capodiferro, M.R., Formenti, G., Caprioli, M., et al. (2022). The mitogenome relationships and phylogeography of barn swallows (*Hirundo rustica*). Mol. Biol. Evol. *39*, msac113. https://doi.org/10.1093/molbev/msac113.
4. Johnston, R.F. (2001). Synanthropic birds of North America. In Avian Ecology and Conservation in an Urbanizing World, J.M. Marzluff, R. Bowman, and R. Donnelly, eds. (Springer US), pp. 49–67. https://doi.org/10.1007/978-1-4615-1531-9_3.
5. Krajcarz, M., Krajcarz, M.T., Baca, M., Baumann, C., Van Neer, W., Popović, D., Sudoł-Procyk, M., Wach, B., Wilczyński, J., Wojenka, M., et al. (2020). Ancestors of domestic cats in Neolithic Central Europe: Isotopic evidence of a synanthropic diet. Proc. Natl. Acad. Sci. USA *117*, 17710–17719. https://doi.org/10.1073/pnas.1918884117.
6. Turner, A. (2010). The Barn Swallow (Poyser). https://doi.org/10.5040/9781472596888.
7. Saino, N., Ambrosini, R., Albetti, B., Caprioli, M., De Giorgio, B., Gatti, E., Liechti, F., Parolini, M., Romano, A., Romano, M., et al. (2017). Migration phenology and breeding success are predicted by methylation of a photoperiodic gene in the barn swallow. Sci. Rep. *7*, 45412. https://doi.org/10.1038/srep45412.
8. Saino, N., Ambrosini, R., Caprioli, M., Liechti, F., Romano, A., Rubolini, D., et al. (2017). Wing morphology, winter ecology, and fecundity selection: evidence for sex-dependence in barn swallows (*Hirundo rustica*). Oecologia *184*, 799–812. https://doi.org/10.1007/s00442-017-3918-0.
9. Saino, N., Romano, M., Rubolini, D., Ambrosini, R., Romano, A., Caprioli, M., Costanzo, A., and Bazzi, G. (2014). A trade-off between reproduction and feather growth in the barn swallow (Hirundo rustica). PLoS One *9*, e96428. https://doi.org/10.1371/journal.pone.0096428.
10. Pap, P.L., Osváth, G., Aparicio, J.M., Bărbos, L., Matyjasiak, P., Rubolini, D., et al. (2015). Sexual dimorphism and population differences in structural properties of barn swallow (*Hirundo rustica*) wing and tail feathers. PLoS One *10*, e0130844. https://doi.org/10.1371/journal.pone.0130844.
11. Pap, P.L., Fülöp, A., Adamkova, M., Cepak, J., Michalkova, R., Safran, R.J., Stermin, A.N., Tomasek, O., Vágási, C.I., Vincze, O., et al. (2019). Selection on multiple sexual signals in two Central and Eastern European populations of the barn swallow. Ecol. Evol. *9*, 11277–11287. https://doi.org/10.1002/ece3.5629.
12. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737–746. https://doi.org/10.1038/s41586-021-03451-0.
13. Garg, V., Dudchenko, O., Wang, J., Khan, A.W., Gupta, S., Kaur, P., Han, K., Saxena, R.K., Kale, S.M., Pham, M., et al. (2022). Chromosome-length genome assemblies of six legume species provide insights into genome organization, evolution, and agronomic traits for crop improvement. J. Adv. Res. *42*, 315–329. https://doi.org/10.1016/j.jare.2021.10.009.
14. Safran, R.J., Scordato, E.S.C., Wilkins, M.R., Hubbard, J.K., Jenkins, B.R., Albrecht, T., Flaxman, S.M., Karaardıç, H., Vortman, Y., Lotem, A., et al. (2016). Genome-wide differentiation in closely related populations: the roles of selection and geographic isolation. Mol. Ecol. *25*, 3865–3883. https://doi.org/10.1111/mec.13740.
15. Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A.H., Xie, D., Chen, G., Guo, C., Faircloth, B.C., et al. (2020). Dense sampling of bird diversity increases power of comparative genomics. Nature *587*, 252–257. https://doi.org/10.1038/s41586-020-2873-9.
16. Formenti, G., Chiara, M., Poveda, L., Francoijs, K.-J., Bonisoli-Alquati, A., Canova, L., Gianfranceschi, L., Horner, D.S., and Saino, N. (2019). SMRT long reads and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow

(*Hirundo rustica rustica*). Gigascience *8*, giy142. https://doi.org/10.1093/gigascience/giy142.

17. Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C., and DePristo, M.A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genom. *13*, 375. https://doi.org/10.1186/1471-2164-13-375.
18. Howe, K., and Wood, J.M.D. (2015). Using optical mapping data for the improvement of vertebrate genome assemblies. Gigascience *4*, 10. https://doi.org/10.1186/s13742-015-0052-y.
19. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2022). A draft human pangenome reference. Preprint at bioRxiv. https://doi.org/10.1101/2022.07.09.499321.
20. Garg, S., Balboa, R., and Kuja, J. (2022). Chromosome-scale haplotype-resolved pangenomics. Trends Genet. *38*, 1103–1107. https://doi.org/10.1016/j.tig.2022.06.011.
21. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. *11*, 1432. https://doi.org/10.1038/s41467-020-14998-3.
22. Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. *35*, 543–548. https://doi.org/10.1093/molbev/msx319.
23. Tomaszkiewicz, M., Medvedev, P., and Makova, K.D. (2017). Y and W Chromosome assemblies: approaches and discoveries. Trends Genet. *33*, 266–282. https://doi.org/10.1016/j.tig.2017.01.008.
24. Malinovskaya, L.P., Tishakova, K., Shnaider, E.P., Borodin, P.M., and Torgasheva, A.A. (2020). Heterochiasmy and sexual dimorphism: the case of the barn swallow (*Hirundo rustica*, Hirundinidae, aves). Genes *11*, 1119. https://doi.org/10.3390/genes11101119.
25. Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. *21*, 245. https://doi.org/10.1186/s13059-020-02134-9.
26. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
27. Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. *42*, D756–D763. https://doi.org/10.1093/nar/gkt1114.
28. Francis, W.R., and Wörheide, G. (2017). Similar ratios of introns to intergenic sequence across animal genomes. Genome Biol. Evol. *9*, 1582–1598. https://doi.org/10.1093/gbe/evx103.
29. Burt, D.W. (2002). Origin and evolution of avian microchromosomes. Cytogenet. Genome Res. *96*, 97–112. https://doi.org/10.1159/000063018.
30. Kim, J., Lee, C., Ko, B.J., Yoo, D.A., Won, S., Phillippy, A., Fedrigo, A., Zhang, G., Howe, K., Wood, J., et al. (2021). False gene and chromosome losses affected by assembly and sequence errors. Preprint at bioRxiv. https://doi.org/10.1101/2021.04.09.438906.
31. Korlach, J., Gedman, G., Kingan, S.B., Chin, C.-S., Howard, J.T., Audet, J.-N., Cantin, L., and Jarvis, E.D. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. Gigascience *6*, 1–16. https://doi.org/10.1093/gigascience/gix085.
32. Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I.T., Novak, A.M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., et al. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature *587*, 246–251. https://doi.org/10.1038/s41586-020-2871-y.
33. Hubisz, M.J., Pollard, K.S., and Siepel, A. (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. Brief. Bioinform. *12*, 41–51. https://doi.org/10.1093/bib/bbq072.
34. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
35. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050. https://doi.org/10.1101/gr.3715005.
36. Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., Storz, J.F., Antunes, A., Greenwold, M.J., Meredith, R.W., et al. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. Science *346*, 1311–1320. https://doi.org/10.1126/science.1251385.
37. Craig, R.J., Suh, A., Wang, M., and Ellegren, H. (2018). Natural selection beyond genes: identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (*Ficedula albicollis*). Mol. Ecol. *27*, 476–492. https://doi.org/10.1111/mec.14462.
38. Axelsson, E., Webster, M.T., Smith, N.G.C., Burt, D.W., and Ellegren, H. (2005). Comparison of the chicken and Turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. Genome Res. *15*, 120–125. https://doi.org/10.1101/gr.3021305.
39. Schield, D.R., Scordato, E.S.C., Smith, C.C.R., Carter, J.K., Cherkaoui, S.I., Gombobaatar, S., Hajib, S., Hanane, S., Hund, A.K., Koyama, K., et al. (2021). Sex-linked genetic diversity and differentiation in a globally distributed avian species complex. Mol. Ecol. *30*, 2313–2332. https://doi.org/10.1111/mec.15885.
40. von Rönn, J.A.C., Shafer, A.B.A., Wolf, J.B.W., Hybridization, G.O.F., von Rönn, J.A.C., Shafer, A.B.A., Wolf, J.B.W., and Hybridization, G.O.F. (2016). Disruptive selection without genome-wide evolution across a migratory divide. Mol. Ecol. *25*, 2529–2541. https://doi.org/10.1111/mec.13521.
41. Scordato, E.S.C., Wilkins, M.R., Semenov, G., Rubtsov, A.S., Kane, N.C., and Safran, R.J. (2017). Genomic variation across two barn swallow hybrid zones reveals traits associated with divergence in sympatry and allopatry. Mol. Ecol. *26*, 5676–5691. https://doi.org/10.1111/mec.14276.
42. Smeds, L., Warmuth, V., Bolivar, P., Uebbing, S., Burri, R., Suh, A., Nater, A., Bureš, S., Garamszegi, L.Z., Hogner, S., et al. (2015). Evolutionary analysis of the female-specific avian W chromosome. Nat. Commun. *6*, 7330. https://doi.org/10.1038/ncomms8330.
43. Murray, G.G.R., Soares, A.E.R., Novak, B.J., Schaefer, N.K., Cahill, J.A., Baker, A.J., Demboski, J.R., Doll, A., Da Fonseca, R.R., Fulton, T.L., et al. (2017). Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science *358*, 951–954. https://doi.org/10.1126/science.aao0960.
44. Corcoran, P., Gossmann, T.I., Barton, H.J., Great Tit HapMap Consortium; Slate, J., and Zeng, K. (2017). Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. Genome Biol. Evol. *9*, 2987–3007. https://doi.org/10.1093/gbe/evx213.
45. El Hou, A., Rocha, D., Venot, E., Blanquet, V., and Philippe, R. (2021). Long-range linkage disequilibrium in French beef cattle breeds. Genet. Sel. Evol. *53*, 63. https://doi.org/10.1186/s12711-021-00657-8.
46. Slatkin, M. (2008). Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. *9*, 477–485. https://doi.org/10.1038/nrg2361.
47. Joiret, M., Mahachie John, J.M., Gusareva, E.S., and Van Steen, K. (2019). Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. BioData Min. *12*, 11. https://doi.org/10.1186/s13040-019-0199-7.

48. Liu, S., He, S., Chen, L., Li, W., Di, J., and Liu, M. (2017). Estimates of linkage disequilibrium and effective population sizes in Chinese Merino (Xinjiang type) sheep by genome-wide SNPs. Genes Genomics *39*, 733–745. https://doi.org/10.1007/s13258-017-0539-2.

49. Pritchard, J.K., Stephens, M., Rosenberg, N.A., and Donnelly, P. (2000). Association mapping in structured populations. Am. J. Hum. Genet. *67*, 170–181. https://doi.org/10.1086/302959.

50. Stapley, J., Birkhead, T.R., Burke, T., and Slate, J. (2010). Pronounced inter- and intrachromosomal variation in linkage disequilibrium across the zebra finch genome. Genome Res. *20*, 496–502. https://doi.org/10.1101/gr.102095.109.

51. Kapusta, A., and Suh, A. (2017). Evolution of bird genomes-a transposon's-eye view. Ann. N. Y. Acad. Sci. *1389*, 164–185. https://doi.org/10.1111/nyas.13295.

52. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

53. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. Brief. Bioinform. *19*, 118–135. https://doi.org/10.1093/bib/bbw089.

54. Sherman, R.M., and Salzberg, S.L. (2020). Pan-genomics in the human genome era. Nat. Rev. Genet. *21*, 243–254. https://doi.org/10.1038/s41576-020-0210-7.

55. Baran, N., Lapidot, A., and Manor, H. (1991). Formation of DNA triplexes accounts for arrests of DNA synthesis at d(TC)n and d(GA)n tracts. Proc. Natl. Acad. Sci. USA *88*, 507–511. https://doi.org/10.1073/pnas.88.2.507.

56. Samadashwily, G.M., Dayn, A., and Mirkin, S.M. (1993). Suicidal nucleotide sequences for DNA polymerization. EMBO J. *12*, 4975–4983. https://doi.org/10.1002/j.1460-2075.1993.tb06191.x.

57. Mirkin, S.M., and Frank-Kamenetskii, M.D. (1994). H-DNA and related structures. Annu. Rev. Biophys. Biomol. Struct. *23*, 541–576. https://doi.org/10.1146/annurev.bb.23.060194.002545.

58. Sirén, J., Garrison, E., Novak, A.M., Paten, B., and Durbin, R. (2020). Haplotype-aware graph indexes. Bioinformatics *36*, 400–407. https://doi.org/10.1093/bioinformatics/btz575.

59. Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. *36*, 875–879. https://doi.org/10.1038/nbt.4227.

60. Papale, L.A., Madrid, A., Li, S., and Alisch, R.S. (2017). Early-life stress links 5-hydroxymethylcytosine to anxiety-related behaviors. Epigenetics *12*, 264–276. https://doi.org/10.1080/15592294.2017.1285986.

61. Vigil, F.A., Mizuno, K., Lucchesi, W., Valls-Comamala, V., and Giese, K.P. (2017). Prevention of long-term memory loss after retrieval by an endogenous CaMKII inhibitor. Sci. Rep. *7*, 4040. https://doi.org/10.1038/s41598-017-04355-8.

62. Notaras, M., and van den Buuse, M. (2020). Neurobiology of BDNF in fear memory, sensitivity to stress, and stress-related disorders. Mol. Psychiatry *25*, 2251–2274. https://doi.org/10.1038/s41380-019-0639-2.

63. O'Rourke, T., Martins, P.T., Asano, R., Tachibana, R.O., Okanoya, K., and Boeckx, C. (2021). Capturing the effects of domestication on vocal learning complexity. Trends Cogn. Sci. *25*, 462–474. https://doi.org/10.1016/j.tics.2021.05.002.

64. Boehler, N.A., Fung, S.W., Hegazi, S., Cheng, A.H., and Cheng, H.-Y.M. (2021). Sox2 ablation in the suprachiasmatic nucleus perturbs anxiety- and depressive-like behaviors. Neurol. Int. *13*, 541–554. https://doi.org/10.3390/neurolint13040054.

65. Wang, H., Sawai, A., Toji, N., Sugioka, R., Shibata, Y., Suzuki, Y., Ji, Y., Hayase, S., Akama, S., Sese, J., et al. (2019). Transcriptional regulatory divergence underpinning species-specific learned vocalization in songbirds. PLoS Biol. *17*, e3000476. https://doi.org/10.1371/journal.pbio.3000476.

66. Kawakami, T., Mugal, C.F., Suh, A., Nater, A., Burri, R., Smeds, L., and Ellegren, H. (2017). Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. Mol. Ecol. *26*, 4158–4172. https://doi.org/10.1111/mec.14197.

67. O'Reilly, P.F., Birney, E., and Balding, D.J. (2008). Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res. *18*, 1304–1313. https://doi.org/10.1101/gr.067181.107.

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

69. Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. *45*, e18. https://doi.org/10.1093/nar/gkw955.

70. Donath, A., Jühling, F., Al-Arab, M., Bernhart, S.H., Reinhardt, F., Stadler, P.F., Middendorf, M., and Bernt, M. (2019). Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. Nucleic Acids Res. *47*, 10543–10552. https://doi.org/10.1093/nar/gkz833.

71. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

72. Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569. https://doi.org/10.1038/nmeth.2474.

73. Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13*, 1050–1054. https://doi.org/10.1038/nmeth.4035.

74. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics *36*, 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025.

75. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.

76. Ghurye, J., Rhie, A., Walenz, B.P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A.M., and Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput. Biol. *15*, e1007273. https://doi.org/10.1371/journal.pcbi.1007273.

77. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. Preprint at arXiv. https://doi.org/10.48550/arXiv.1207.3907.

78. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). the sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

79. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

80. Chow, W., Brugger, K., Caccamo, M., Sealy, I., Torrance, J., and Howe, K. (2016). gEVAL - a web-based browser for evaluating genome assemblies. Bioinformatics *32*, 2508–2510. https://doi.org/10.1093/bioinformatics/btw159.

81. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

82. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. *5*, R12. https://doi.org/10.1186/gb-2004-5-2-r12.

83. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118. https://doi.org/10.1371/journal.pcbi.1003118.

84. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868. https://doi.org/10.1093/bioinformatics/btx699.

85. Morgulis, A., Gertz, E.M., Schäffer, A.A., and Agarwala, R. (2006). WindowMasker: window-based masker for sequenced genomes. Bioinformatics *22*, 134–141. https://doi.org/10.1093/bioinformatics/bti774.

86. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics *Chapter 4*, 4.10.1–4.10.14. https://doi.org/10.1002/0471250953.bi0410s25.

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

88. Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. *34*, 1812–1819. https://doi.org/10.1093/molbev/msx116.

89. Hickey, G., Paten, B., Earl, D., Zerbino, D., and Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics *29*, 1341–1342. https://doi.org/10.1093/bioinformatics/btt128.

90. Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D., and Woolf, P.J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinf. *10*, 161. https://doi.org/10.1186/1471-2105-10-161.

91. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440. https://doi.org/10.1093/bioinformatics/bti525.

92. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. *35*, 1547–1549. https://doi.org/10.1093/molbev/msy096.

93. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

94. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

95. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner. Berkeley, CA (United States): Lawrence Berkeley National Lab.(LBNL); 2014. https://www.osti.gov/servlets/purl/1241166.

96. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

97. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192. https://doi.org/10.1093/bib/bbs017.

98. Gel, B., and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics *33*, 3088–3090. https://doi.org/10.1093/bioinformatics/btx346.

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

100. Dong, S.-S., He, W.-M., Ji, J.-J., Zhang, C., Guo, Y., and Yang, T.-L. (2021). LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Brief. Bioinform. *22*, bbaa227. https://doi.org/10.1093/bib/bbaa227.

101. Fan, Z., Yue, B., Zhang, X., Du, L., and Jian, Z. (2017). CpGIScan: an ultrafast tool for CpG islands identification from genome sequence. Curr. Bioinform. *12*, 181–184. https://doi.org/10.2174/1574893611666160907111325.

102. Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G.W., Schöenhuth, A., and Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. Preprint at bioRxiv. https://doi.org/10.1101/085050.

103. Gautier, M., and Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics *28*, 1176–1177. https://doi.org/10.1093/bioinformatics/bts115.

104. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. *36*, 983–987. https://doi.org/10.1038/nbt.4235.

105. Yun, T., Li, H., Chang, P.-C., Lin, M.F., Carroll, A., and McLean, C.Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics *36*, 5582–5589. https://doi.org/10.1093/bioinformatics/btaa1081.

106. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. *37*, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9.

107. Hall, M. (2022). Rasusa: randomly subsample sequencing reads to a specified coverage. J. Open Source Softw. *7*, 3941. https://doi.org/10.21105/joss.03941.

108. Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. Genome Biol. *21*, 265. https://doi.org/10.1186/s13059-020-02168-z.

109. Zhang, X., Kaplow, I.M., Wirthlin, M., Park, T.Y., and Pfenning, A.R. (2020). HALPER facilitates the identification of regulatory element orthologs across species. Bioinformatics *36*, 4339–4340. https://doi.org/10.1093/bioinformatics/btaa493.

110. Wickham, H. (2021). ggplot2: Elegant Graphics for Data Analysis 2016. https//ggplot2. tidyverse.org.

111. Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). Circlize Implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812. https://doi.org/10.1093/bioinformatics/btu393.

112. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

113. Beyer, W., Novak, A.M., Hickey, G., Chan, J., Tan, V., Paten, B., and Zerbino, D.R. (2019). Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics *35*, 5318–5320. https://doi.org/10.1093/bioinformatics/btz597.

114. Challis, R., Richards, E., Rajan, J., Cochrane, G., and Blaxter, M. (2020). BlobToolKit–Interactive quality assessment of genome assemblies. G3 *10*, 1361–1374. https://doi.org/10.1534/g3.119.400908.

115. Cabanettes, F., and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ *6*, e4958. https://doi.org/10.7717/peerj.4958.

116. Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., and Liu, X. (2021). rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide

association study. Dev. Reprod. Biol. *19*, 619–628. https://doi.org/10.1016/j.gpb.2020.10.007.

117. R Core Team (2020). R: A Language and Environment for Statistical Computing.

118. Griffiths, R., Daan, S., and Dijkstra, C. (1996). Sex identification in birds using two CHD genes. Proc. Biol. Sci. *263*, 1251–1256. https://doi.org/10.1098/rspb.1996.0184.

119. Stanyon, R., and Galleni, L. (1991). A rapid fibroblast culture technique for high resolution karyotypes. Bolletino di zoologia *58*, 81–83. https://doi.org/10.1080/11250009109355732.

120. Formenti, G., Rhie, A., Balacco, J., Haase, B., Mountcastle, J., Fedrigo, O., Brown, S., Capodiferro, M.R., Al-Ajli, F.O., Ambrosini, R., et al. (2021). Complete vertebrate mitogenomes reveal widespread gene duplications and repeats. Genome Biol. *22*, 120. https://doi.org/10.1186/s13059-021-02336-9.

121. Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D.-L., Sims, Y., Torrance, J., Tracey, A., and Wood, J. (2021). Significantly improving the quality of genome assemblies through curation. Gigascience *10*, giaa153. https://doi.org/10.1093/gigascience/giaa153.

122. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: assessing genome assembly and annotation completeness. Methods Mol. Biol. *1962*, 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14.

123. Kuhl, H., Frankl-Vilches, C., Bakker, A., Mayr, G., Nikolaus, G., Boerno, S.T., et al. (2021). An unbiased molecular approach using 3′-UTRs resolves the avian family-level tree of life. Mol. Biol. *38*, 108–127. https://doi.org/10.1093/molbev/msaa191.

124. Zar, J.H. (1972). Significance testing of the spearman rank correlation coefficient. J. Am. Stat. Assoc. *67*, 578–580. https://doi.org/10.1080/01621459.1972.10481251.

125. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*, 695–716. https://doi.org/10.1038/nature03154.

126. Smit AFA, Hubley R, Green P. RepeatMasker. http://www.repeatmasker.org

127. Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11. https://doi.org/10.1186/s13100-015-0041-9.

128. Mann, H.B., and Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist. *18*, 50–60. https://doi.org/10.1214/aoms/1177730491.

129. Siepel, A. PhastCons HOWTO. http://compgen.cshl.edu/phast/phastCons-HOWTO.html

130. Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin *1*, 80–83. http//www. jstor.org/stable/3001968.

131. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482. https://doi.org/10.1038/nature10530.

132. Chang, C.C. (2020). Data management and summary statistics with PLINK. Methods Mol. Biol. *2090*, 49–65. https://doi.org/10.1007/978-1-0716-0199-0_3.

133. Ardlie, K.G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. *3*, 299–309. https://doi.org/10.1038/nrg777.

134. Bejarano, D., Martínez, R., Manrique, C., Parra, L.M., Rocha, J.F., Gómez, Y., Abuabara, Y., and Gallego, J. (2018). Linkage disequilibrium levels and allele frequency distribution in Blanco Orejinegro and Romosinuano Creole cattle using medium density SNP chip data. Genet. Mol. Biol. *41*, 426–433. https://doi.org/10.1590/1678-4685-GMB-2016-0310.

135. Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J.W., Landolin, J.M., Maurer, N., Kudrna, D., Hardigan, M.A., Steiner, C.C., et al. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. Sci. Data *7*, 399. https://doi.org/10.1038/s41597-020-00743-4.

136. Nurk, S., Walenz, B.P., Rhie, A., Vollger, M.R., Logsdon, G.A., Grothe, R., Miga, K.H., Eichler, E.E., Phillippy, A.M., and Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. *30*, 1291–1305. https://doi.org/10.1101/gr.263566.120.

137. Mc Cartney, A.M., Shafin, K., Alonge, M., Bzikadze, A.V., Formenti, G., Fungtammasan, A., Howe, K., Jain, C., Koren, S., Logsdon, G.A., et al. (2022). Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nat. Methods *19*, 687–695. https://doi.org/10.1038/s41592-022-01440-3.

138. Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. https://doi.org/10.1080/14786440009463897.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Proteinase K	VWR	Cat#1.24568.0100
RNAse A	N/A	N/A
Critical commercial assays		
Bionano animal tissue DNA isolation fibrous tissue protocol	Bionano genomics	cat# RE-013-10
Circulomics Nanobind Tissue Big DNA kit	Circulomics (now Pacific Biosciences)	SKU NB-900-701-01 (Not commercialized anymore)
Genome Library Kit	10x Genomics Chromium	v2 PN-120258
Gel Bead Kit	10x Genomics Chromium	v2 PN-120258
Genome Chip Kit	10x Genomics Chromium	v2 PN-120257
i7 Multiplex Kit	10x Genomics Chromium	PN-120262
Arima-HiC kit	Arima Genomics	P/N: A510008
KAPA Hyper Prep kit	Roche	P/N: KK8504
QIAGEN RNAeasy kit	QIAGEN	cat# 74104
Qubit™ RNA BR Assay Kit	ThermoFisher Scientific	cat# Q10210
NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification Module	New England BioLabs	cat# E6421S
Iso-Seq Express Oligo Kit	Pacific Biosciences	PN 10 1-737-500
ProNex® Beads	Promega	Cat# NG2001
SMRTbell Express Template Prep Kit 2.0	Pacific Biosciences	PN 101-685-400; PN: 100-938-900
Iso-seq sequencing kit 3.0	Pacific Biosciences	#101-597-800
TruSeq Stranded mRNA LT Sample Prep Kit	Illumina	N/A
QIAGEN Genomic-tip	Qiagen	cat# 10223
Deposited data		
de novo assembly for *Hirundo rustica*	This study	RefSeq: GCF_015227805.1. Genbank: GCA_015227805.3, GCA_015227815.3. NCBI BioProject: PRJNA909772
10x and Hi-C genomic data for bHirRus1 reference assembly	This study	SRA: SRR22566724, SRR22566725, SRR22566726, SRR22566727 (10x). SRA: SRR22566728, SRR22566729 (Hi-C).
PacBio CLR reads and Bionano DLS optical maps for bHirRus1 reference assembly	reused from Formenti et al.[16]	SRA: SRR7589801 and SRR7589802 (PacBio CLR reads). Bionano optical maps are available in the *GigaScience* GigaDB repository associated to Formenti et al.[16]
Hifi sequencing reads	This study	SRA: SRR22588214, SRR22588215, SRR22588216, SRR22588217, SRR2258821.
Isoseq data	This study	SRA: SRR9184408 and SRR9184409.
RNAseq data	This study	SRA: SRR13516425, SRR13516426, SRR13516427, and SRR10853074.
Raw fastq reads	Safran et al.[14]	NCBI BioProject: PRJNA323498.
Raw fastq reads	von Rönn et al.[40]	NCBI BioProject: PRJNA296600.
Raw fastq reads	Scordato et al.[41]	NCBI BioProject: PRJNA323498.
Raw fastq reads	Smith et al.[2]	NCBI BioProject: PRJNA323498.
Raw fastq reads	Schield et al.[39]	NCBI BioProject: PRJNA323498.
Newly generated genomic resources (variants catalog, pangenome, Cactus alignment)	This study	Dataverse: https://doi.org/10.13130/RD_UNIMI/IDALZG
Hirundo r. rustica mitochondrial Reference Sequence	Lombardo et al.[3]	GenBank: MZ905359

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Cell lines		
Barn swallow cells cultured for karyotype reconstruction	This study	N/A
Software and algorithms		
All scripts written and used for this study	This study	https://doi.org/10.5281/zenodo.7474288
VGP genome assembly pipeline 1.6	Rhie et al.[12]	https://vertebrategenomesproject.org/
bowtie2 v2.4.1	Langmead and Salzberg[68]	https://github.com/BenLangmead/bowtie2
NOVOplasty	Dierckxsens et al.[69]	https://github.com/ndierckx/NOVOPlasty
MITOS2	Donath et al.[70]	http://mitos2.bioinf.uni-leipzig.de/index.py
Genomescope2.0	Ranallo-Benavidez et al.[21]	http://qb.cshl.edu/genomescope/genomescope2.0/
Meryl	Rhie et al.[25]	https://github.com/marbl/meryl
Mash	Ondov et al.[71]	https://github.com/marbl/mash
process_10xReads.py script	ucdavis-bioinformatics	https://github.com/ucdavis-bioinformatics/proc10xG
FALCON	Chin et al.[72]	https://pb-falcon.readthedocs.io/en/latest/
FALCON-unzip	Chin et al.[73]	https://pb-falcon.readthedocs.io/en/latest/about.html
Arrow	Chin et al.[72]	N/A
Purge_dups	Guan et al.[74]	https://github.com/dfguan/purge_dups
Merqury	Rhie et al.[25]	https://github.com/marbl/merqury
scaff10X v2.0–2.1	N/A	https://github.com/wtsi-hpag/Scaff10X
Bionano Solve v3.2.1	Bionano genomics	https://bionanogenomics.com/support/software-downloads/
Arima Genomics mapping pipeline	Arima genomics	https://github.com/ArimaGenomics/mapping_pipeline
BWA-MEM v0.7.17-r1188	Li and Durbin[75]	https://github.com/lh3/bwa
Salsa v2.2	Ghurye et al.[76]	https://github.com/marbl/SALSA
Longranger align v2.2.2	10x Genomics	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
Freebayes v1.2.0, v1.3.1	Garrison and Marth[77]	https://github.com/freebayes/freebayes
bcftools v1.1	Li et al.[78]; Danecek et al.[79]	https://samtools.github.io/bcftools/
genome evaluation browser gEVAL	Chow et al.[80]	geval.org.uk
BUSCO v4.1.4	Simão et al.[26]	https://gitlab.com/ezlab/busco
BLAST 2.10.1+	Camacho et al.[81]	The latest version of BLAST can be retrieved from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST
MUMMer NUCmer	Kurtz et al.[82]	https://mummer.sourceforge.net/
NCBI Eukaryotic genome annotation pipeline	Pruitt et al.[27]	https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/
GenomicFeatures	Lawrence et al.[83]	https://bioconductor.org/packages/release/bioc/html/GenomicFeatures.html
chromosome_size software	N/A	https://git.mpi-cbg.de/dibrov/chromosome_size#citation
samtools v1.9, v1.10	Li et al.[78]; Danecek et al.[79]	https://github.com/samtools/
mosdepth	Pedersen and Quinlan[84]	https://github.com/brentp/mosdepth
PretextMap	N/A	https://github.com/wtsi-hpag/PretextMap
PretextView	N/A	https://github.com/wtsi-hpag/PretextView

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
WindowMasker v1.0.0	Morgulis et al.[85]	WM is included in the NCBI C++ toolkit. The source code for the entire toolkit is available at ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools++/CURRENT/.
RepeatMasker v4.1.0	Tarailo-Graovac and Chen[86]	http://www.repeatmasker.org
bedtools v2.29.2	Quinlan and Hall[87]	https://github.com/arq5x/bedtools2
Cactus v1.3.0	Armstrong et al.[32]	https://github.com/ComparativeGenomicsToolkit/cactus
TimeTree	Kumar et al.[88]	http://www.timetree.org/
HAL toolkit	Hickey et al.[89]	http://github.com/glennhickey/hal
PHAST v1.5	Hubisz et al.[33]	http://compgen.bscb.cornell.edu/phast
maf_stream	N/A	https://github.com/joelarmstrong/maf_stream
msa_view	Hubisz et al.[33]	http://compgen.cshl.edu/phast/
phyloFit	Hubisz et al.[33]	http://compgen.cshl.edu/phast/
PhyloP	Hubisz et al.[33]	http://compgen.cshl.edu/phast/
PhastCons	Hubisz et al.[33]	http://compgen.cshl.edu/phast/
phyloBoot	Hubisz et al.[33]	http://compgen.cshl.edu/phast/
consEntropy	Hubisz et al.[33]	http://compgen.cshl.edu/phast/
gage R package	Luo et al.[90]	https://bioconductor.org/packages/release/bioc/html/gage.html
bioMart R package	Durinck et al.[91]	https://bioconductor.org/packages/release/bioc/html/biomaRt.html
MEGA	Kumar et al.[92]	https://www.megasoftware.net/
SRA Toolkit v2.9.1	N/A	https://github.com/ncbi/sra-tools
Fastqc v0.11.9	N/A	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Multiqc v1.9	Ewels et al.[93]	https://github.com/ewels/MultiQC
Cutadapt v2.10, v3.2	Martin[94]	https://cutadapt.readthedocs.io/en/stable/installation.html
BBMap v38.18	Bushnell[95]	https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbmap-guide/
Picard MarkDuplicates v2.23.4	N/A	https://broadinstitute.github.io/picard/
Bam clipOverlap v1.0.14	N/A	https://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap
VGP assembly pipeline freebayes-polish script	Rhie et al.[12]	https://github.com/VGP/vgp-assembly/blob/master/pipeline/freebayes-polish/freebayes_v1.3.sh
Script generating masked ranges within a fasta file	N/A	https://gist.github.com/danielecook/cfaa5c359d99bcad3200
VCFtools v.0.1.16	Danecek et al.[96]	https://github.com/vcftools/vcftools
Integrative Genomics Viewer (IGV)	Thorvaldsdóttir et al.[97]	https://software.broadinstitute.org/software/igv/
karyoploteR R package	Gel and Serra[98]	https://bioconductor.org/packages/devel/bioc/vignettes/karyoploteR/inst/doc/karyoploteR.html
Plink v1.9	Purcell et al.[99]	https://zzz.bwh.harvard.edu/plink/index.shtml
LDBlockShow v1.36	Dong et al.[100]	https://github.com/BGI-shenzhen/LDBlockShow
cpgiscan v1.0	Fan et al.[101]	https://github.com/jzuoyi/cpgiscan
WhatsHap v0.18; WhatsHap development version v.1.2.dev2+g3dffe4a	Martin et al.[102]	https://github.com/whatshap/whatshap

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Rehh R package	Gautier and Vitalis[103]	https://cran.r-project.org/web/packages/rehh/index.html
qvalue R package	N/A	https://github.com/StoreyLab/qvalue
pbmm2 v1.3.0, v1.4.0	N/A	https://github.com/PacificBiosciences/pbmm2
DeepVariant v1.0.0	Poplin et al.[104]	https://github.com/google/deepvariant
GLNexus pipeline for HiFi joint calling	Yun et al.[105]	https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake
pbsv v2.6.0	Wenger et al.[106]	https://github.com/PacificBiosciences/pbsv
Rasusa v0.3.0	Hall[107]	https://github.com/mbhall88/rasusa
Hifiasm v0.13-r307	Cheng et al.[52]	https://github.com/chhylp123/hifiasm
Cactus Pangenome Pipeline	Armstrong et al.[32]	https://github.com/ComparativeGenomicsToolkit/cactus/blob/master/doc/pangenome.md
Minigraph v0.14-r415	Li et al.[108]	https://github.com/lh3/minigraph
HALPER	Zhang et al.[109]	https://github.com/pfenninglab/halLiftover-postprocessing
ggplot2 R package	Wickham[110]	https://github.com/tidyverse/ggplot2
Circlize	Gu et al.[111]	https://github.com/jokergoo/circlize
ComplexHeatmap	Gu et al.[112]	https://github.com/jokergoo/ComplexHeatmap
SequenceTubeMap	Beyer et al.[113]	https://github.com/vgteam/sequenceTubeMap
BloobToolKit	Challis et al.[114]	https://blobtoolkit.genomehubs.org/
D-genies	Cabanettes and Klopp[115]	https://dgenies.toulouse.inra.fr/
CMplot	Yin[116]	https://github.com/YinLiLin/CMplot
asm_stats (VGP genome assembly pipeline 1.6)	Rhie et al.[12]	https://github.com/VGP/vgp-assembly/blob/master/pipeline/stats/asm_stats.sh
R studio	R core team[117]	https://cran.r-project.org/
Variation graph toolkit	Garrison et al.[59]	https://github.com/vgteam/vg

RESOURCE AVAILABILITY

Lead contact

Further information about datasets, protocols, and workflows used should be directed to and will be fulfilled by the lead contact, Giulio Formenti (gformenti@rockefeller.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Primary and alternate assemblies (bHirRus1) presented in this study are available on NCBI. All raw data supporting the genome assembly are available in Genbank and also on GenomeArk (https://vgp.github.io/genomeark/Hirundo_rustica/). Additional HiFi sequencing data used to generate the pangenome, IsoSeq, and RNAseq data used for annotation are available in Genbank. All accession numbers are listed in the key resources table. Newly generated genomic resources (SNP catalog, Cactus alignment, and pangenome graph) have been deposited at Dataverse repository (https://dataverse.unimi.it). DOIs are listed in the key resources table. This paper also analyzes existing, publicly available data. The accession numbers for these datasets are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sampling for sequencing

For the de novo genome assembly, tissues were collected from the same ringed barn swallow female whose blood was used for producing the previous barn swallow 'Chelidonia' assembly.[16] The individual was recaptured in June 2018 in the same farm near Milan (45.4N 9.3E) and euthanized under permission N. 5104 issued on 11.04.2018 by Regione Lombardia. Tissues were dissected by an experienced avian veterinary, flash frozen immediately after dissection, and stored at −80°C. The absence of any mistake in sample handling was further corroborated by manual inspection of read alignments of the newly generated reads to the Chelidonia assembly.

For HiFi sequencing, ~100 μL of blood from five Italian barn swallows (*H. r. rustica*), were collected in heparinized capillary tubes through a minimally invasive sampling procedure in June 2019 (sample A1 and A2), July 2020 (sample 2), April 2019 (sample 3) and May 2019 (sample 4). Sampling was performed under permission 3268 of 12.03.2019 by Regione Lombardia. Samples from Matera were collected by Istituto Nazionale per la Protezione e la Ricerca Ambientale (ISPRA) under the authorization of Law 157/1992 [Art.4 (1) and Art. 7 (5)]. Samples from Oleggio (NO) were collected by the Università degli Studi di Milano under the authorization of the Provincia di Novara, Ufficio Caccia e Pesca Acque Interne, D.D. n. 973 (issued on May 15, 2019). Sampling locations are reported in Table S11A.

Karyotype reconstruction

To confirm the chromosomal structure of our assembly, a karyotype for the barn swallow was generated using a cultured cell protocol. Tissue biopsies were obtained from a male *Hirundo r. rustica* sampled under permit N. 3268 issued on 12.03.2019 by Regione Lombardia. The sex of the individual was confirmed by PCR amplification of sex-specific genomic regions as described in Griffith et al., 1996.[118] Cells were cultured in a medium composed of 50% RPMI1640 and 50% Iscove's Modified Dulbecco's Medium, supplemented with 10% fetal bovine serum, 1% penicillin (10,000 units/ml) - streptomycin (10 mg/mL), 1% gentamycin sulfate (10 mg/mL), 0.5% amphotericin B (250 μg/ml) and 1% L-glutamine (200 mM) and incubated at 41°C with 5% CO_2. Chromosome preparations were made following standard procedures.[119] In brief, after 4 h of treatment in 0.01 ng/mL colcemid, the cells are removed by standard trypsination and placed in a 15 mL tube. Cells are then centrifuged at 10,000 g, surnatant is removed and substituted with a 1:1 mixture of 0.075 M KC1 and 0.4% sodium citrate (hypotonic treatment). After a 20-min exposure at 37°C the cells are pelleted by centrifugation and fixed in methanol:acetic acid fixative (at a ratio of 3:1). Slides are then prepared by dropping metaphases with a Pasteur pipette onto a clean glass microscope slide. Diploid number and chromosome morphology were determined from the analyses of 20 mitotic cells stained with DAPI.

METHOD DETAILS

DNA extraction

HMW (High Molecular Weight) DNA was extracted from the muscle tissue of the samples female barn swallow with the Bionano animal tissue DNA isolation fibrous tissue protocol (cat# RE-013-10; document number 30071). Approximately 55 mg of frozen muscle tissue was fixed in formaldehyde (2%) and homogenised with the Qiagen TissueRuptor. The lysate was included in agarose plugs, which were then treated with Proteinase K and RNase A. The DNA was recovered and purified from the plugs through a drop dialysis with 1x TE. Pulsed Field Gel Electrophoresis (PFGE; Pippin Pulse, SAGE Science, Beverly, MA) and Qubit were used for DNA quality control. According to the PFGE run, a large fraction of the isolated DNA was >250kbp.

For HiFi sequencing, High Molecular Weight (HMW) DNA was extracted from whole blood for samples A1 and A2, while for the other HiFi samples (2, 3 and 4) the starting material was centrifuged blood. The Circulomics Nanobind Tissue Big DNA kit (SKU NB-900-701-01) was used to extract HMW DNA, following manufacturer's instructions. DNA absorbance was checked as quality and purity control by Nanodrop and average fragments length was verified with a Pulsed Field Gel Electrophoresis (PFGE). To perform PFGE, the Pulsaphor system with a hexagonal electrode array (Amersham Pharmacia Biotech) was employed. Genomic DNA was loaded on a 1% agarose gel in 0.5X TBE (running conditions: 165V, 60 s pulses for the first 12 h, 90 s pulses for the last 12 h; 8°C). Gel was stained with Ethidium Bromide 2 μg/mL in TBE 0.5X for 30 min; to acquire images, Geldoc (Bio-Rad) was used. To perform a second round of sequencing and achieve a higher coverage, DNA was re-extracted from samples A1,2,3,4 using the Qiagen Genomic tip columns and protocol at a PacBio sequencing service provider at Brigham Young University, Provo, UT (USA).

Library preparation and sequencing

Genomic data from four different sequencing technologies were used for the assembly: Pacific Biosciences (PacBio) CLR long-reads, 10x Genomics linked reads (short-reads), Bionano optical maps with one restriction enzyme (DLS) labeling, and Hi-C reads from Arima Genomics. PacBio long-reads and Bionano optical maps were reused from Chelidonia assembly.[16] Linked-reads libraries were generated using the 10x Genomics Chromium platform (Genome Library Kit & Gel Bead Kit v2 PN-120258, Genome Chip Kit v2 PN-120257, i7 Multiplex Kit PN-120262) and sequenced on an Illumina NovaSeq S4 150bp PE lane at ~60X coverage. Hi-C libraries were generated by Arima Genomics (https://arimagenomics.com/) using muscle *in-vivo* cross-linking with the Arima-HiC kit

(P/N: A510008) with 2-enzymes proximity ligation. Proximally-ligated DNA was subjected to shearing, size-selection (~200–600bp) with SPRI beads, and enrichment with streptavidin beads for the biotin-labelled DNA. KAPA Hyper Prep kit (P/N: KK8504) was employed to generate libraries compatible with Illumina technologies. The libraries were amplified through PCR and purified with SPRI beads. Libraries were sequenced on a Illumina HiSeq X (~60X coverage) after a quality check with Bioanalyzer and qPCR. A quality control for each sequencing data type was performed with Mash[71] to detect potential outlier sequencing runs or species contamination. Mash was run with 21-mers to generate sketches of size 10,000. No contamination was detected.

To generate HiFi data, HMW DNA was sequenced by our PacBio sequencing service provider at Brigham Young University, where it was sheared using a Megaruptor 3. Libraries were prepared using the PacBio "SMRTbell express template Prep kit 2.0". Final size selection was performed using the Blue Pippin.

QUANTIFICATION AND STATISTICAL ANALYSIS

Mitogenome assembly

A *de novo* assembly of the barn swallow mitogenome was generated from 10X reads, which were firstly trimmed with the process_10xReads.py script from proc10xG (https://github.com/ucdavis-bioinformatics/proc10xG) with -a and -b 16 parameters. Trimmed reads were aligned to the Chelidonia assembly[16] with bowtie2[68] and unmapped reads were extracted. NOVOplasty[69] was run with default parameters (read length = 151, insert size = 300) to assemble the mitogenome *de novo* from the unmapped reads. The mitogenome annotation was performed with MITOS2.[70] As sanity check, we aligned and mapped our complete mitochondrial sequence to the *Hirundo r. rustica* mitochondrial Reference Sequence (HrrRS, GenBank accession number MZ905359), which is included in a companion study on barn swallow mitogenome relationships.[3]

Reference genome assembly

Prior to the assembly, Genomescope2.0[21] was used to estimate genome size, heterozygosity and repeat content through statistical analyses of *k*-mer profiles in unassembled sequencing data. Genomescope2.0[21] was run online (http://qb.cshl.edu/genomescope/genomescope2.0/) starting from the *k*-mer (31 bp) histogram generated with Meryl[25] using the 10X linked reads with barcodes (i.e. the first 23 bp of the forward read) trimmed off. Newly generated sequencing data were combined with PacBio CLR long reads and Bionano optical maps already available for the same individual.[16] The assembly was performed on the DNAnexus cloud-based informatic platform for genomic data analyses (https://www.dnanexus.com/) using the VGP standard genome assembly pipeline 1.6[12] (https://github.com/VGP/vgp-assembly; Figure 1A). PacBio subreads from Formenti et al. 2019[16] were used in the first FALCON[72] contigging step. A genome size estimate of 1.31 Gbp (http://www.genomesize.com/) was used for read coverage calculation. Pre-assembled contigs underwent a phasing step with FALCON-unzip[73] (smrtanalysis 3.0.0) and a first round of Arrow[72] (smrtanalysis 5.1.0.26412) polishing. FALCON and FALCON-unzip were run with default parameters, with the exception of parameters related to the identification of read overlaps. Raw reads overlaps were computed with DALIGNER options -k14 -e0.75 -s100 -l2500 -h240 -w8, and pre-assembled reads (preads) overlaps with DALIGNER options -k24 -e.90 -s100 -l1000 -h600. FALCON-unzip generated a set of primary contigs (labeled c1) representing the primary pseudo-haplotype, and a set of alternate haplotigs (c2), representing the secondary haplotypes (Figure 1A). Purge_dups[74] was run on c1 primary contigs to remove any retained haplotig from the primary assembly, particularly in highly divergent regions, and to remove overlaps, collapsed repeats and low- and high-coverage contigs. Purged primary contigs (p1) were scaffolded, whilst all the alternate sequences were included into the p2 intermediate. The latter was merged with c2 alternate haplotigs and subjected to another round of purge_dups to remove additional haplotigs and overlaps. Purged alternate haplotigs (q2) were employed during the polishing step (Figure 1A). To confirm the removal of haplotigs and overlaps, the evaluation tool Merqury[25] was run on primary and alternate contigs before and after purging. After purge_dups, a three-steps scaffolding strategy was performed on the p1 purged primary contigs using Illumina short-reads (10x Genomics), Bionano optical maps and Hi-C reads (Figure 1A). To join proximal contigs, 10x linked reads were aligned to the p1 intermediate in two rounds and an adjacency matrix was produced from the barcodes using scaff10X v2.0–2.1 (https://github.com/wtsi-hpag/Scaff10X). Two scaffolding rounds were performed with options -matrix 2000 -reads 12 -link 10 and then -matrix 2000 -reads 8 -link 10. Contigs were then joined with 100 bp gaps ('N's). The resulting s1 intermediate was then scaffolded with Bionano DLS optical maps[16] using Bionano Solve v3.2.1 in non-haplotype assembly mode with a DLE-1 one enzyme non-nicking approach, obtaining s2. Finally, Hi-C reads from Arima were aligned to the s2 intermediate with the Arima Genomics mapping pipeline (https://github.com/ArimaGenomics/mapping_pipeline). Forward and reverse reads were aligned independently with BWA-MEM[75] with the -B8 parameter and filtered with a minimum mapping quality of 10. Reads containing a restriction enzyme site were trimmed at the 3′ end, and the aligned single reads were paired again. Processed alignments were employed for scaffolding with Salsa v2.2[76] with -m yes -i 5 -p yes parameters and -e GATC, GANTC to indicate restriction enzymes used for library generation. Polishing was performed to improve the assembly per-base accuracy (QV).[12] We targeted Q40 (99.99% accuracy or 1 error/10 kbp).[12] To prevent haplotype switches and overpolishing of NUMTs,[12,120] s3 scaffolded primary assembly was merged with q2 alternate combined haplotigs and the barn swallow mitogenome from NOVOplasty[69] (Figures 1A and S1). The s4 combined intermediate was polished with Arrow (pacific Biosciences; smrtanalysis 5.1.0.26412) with the command 'pbalign –minAccuracy = 0.75 –minLength = 50–minAnchorSize = 12–maxDivergence = 30–concordant–algorithm = blasr –algorithmOptions = –useQuality –maxHits = 1 –hitPolicy = random –seed = 1' for read alignment, and with 'variantCaller –skipUnrecognizedContigs haploid -x 5 -q 20 -X120 -v –algorithm = arrow' for consensus polishing, using PacBio CLR (t1). Two additional rounds of polishing with linked-reads were performed on t1, generating

the t2 intermediate, and the final t3 polished assembly. In this step, raw-reads were aligned with Longranger align 2.2.2 and variants were called with Freebayes v1.2.0[77] with default parameters. Finally, bcftools consensus[78] with options -i 'QUAL>1 && (GT = "AA" || GT = "Aa")' -Hla was used to generate the consensus. The assembly was named 'bHirRus1' after the individual used for sequencing, which in turn is based on VGP guidelines for genome identifiers.[12]

Manual curation

Manual assembly curation entails the removal of contaminants and false duplications, the correction of structural assembly errors and the identification and assignment of chromosomal units. For bHirRus1, a dedicated decontamination pipeline, the genome evaluation browser gEVAL[80] (geval.org.uk) and HiGlass Hi-C 2D maps were used.[121] Since no reference for chromosome assignment was already established for the barn swallow, chromosomes were numbered in decreasing size order. A second curation step was performed using the results from BUSCO 4.1.4,[22,26,122] which indirectly assessed functional completeness through the prediction of highly conserved BUSCO vertebrate genes (complete, complete and single-copy, complete and duplicated, fragmented and missing). The absence, duplication or fragmentation of BUSCO genes can be evidence of assembly errors or missing sequences. BUSCO was run with the vertebrata_odb10 database and 'chicken' as training species for gene prediction on bHirRus1 and Chelidonia to assess differences in functional completeness, but also on the alternate assembly and the assembly pipeline intermediates c1, p1 and p2, to assess whether purge_dups[74] removed unintended sequences from the primary assembly. The BUSCO results were manually evaluated to detect missing genes in bHirRus1 that were found in the other assemblies, and could, therefore, be recovered. Nucleotide-nucleotide BLAST 2.10.1+[81] was used to search in bHirRus1 the sequence of the missing genes retrieved from the corresponding assembly. These genes were erroneously not detected by BUSCO in bHirRus1. To confirm the presence of the genes found with BLAST and rescue the remaining bHirRus1 missing genes from the other assemblies, the scaffold or contig sequences containing the predicted BUSCO genes were aligned to bHirRus1 with MUMMer NUCmer.[82] The alignment files were filtered maintaining only query alignment >1 kbp with an identity >98% with the reference sequence. Alignment coordinates were then manually evaluated. If the gene coordinates in the scaffolds failed to align to bHirRus1, the missing scaffold fragments were extracted from Chelidonia and the alternate assembly and added to bHirRus1. The rescued sequences were trimmed accordingly to avoid the insertion of duplicates and gaps. BUSCO and BLAST analysis were repeated on the new assembly version to confirm the addition of the rescued genes.

Annotation

Total RNA was extracted and purified using the QIAGEN RNAeasy kit (Cat. No. 74104). For each tissue type (brain and ovary), ~30 mg was used, kept on dry ice and cut into 2 mm pieces before being disrupted and homogenised with the Qiagen TissueRuptor II (Cat No./ID: 9,002,755). The RNA quality of all samples was measured using a Fragment Analyzer (Agilent Technologies, Santa Clara, CA) and quantified with a Qubit 2 Fluorometer (Qubit RNA BR Assay Kit - Catalog number: Q10210). PacBio Iso-seq libraries were prepared according to the "Procedure & Checklist – Iso-Seq Express Template Preparation for Sequel and Sequel II Systems (PN 101-763-800 Version 01)". Briefly, cDNA was reverse transcribed using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (New England BioLabs, cat. no. E6421S) and Iso-Seq Express Oligo Kit (PacBio PN 10 1-737-500) from 300 ng of total RNA for both brain and ovary. Amplified cDNA was cleaned with ProNex Beads (Promega - Catalog numbers: NG2001). For each sample, a PacBio library was prepared using the Pacific Biosciences SMRTbell Express Template Prep Kit 2.0 (PN 101-685-400) following the manufacturer protocol. PacBio Iso-seq libraries were sequenced on a PacBio Sequel using sequencing chemistry 3.0 and with 20 h movie time, 4 h pre-extension and PacBio 1M v3 (#101-531-000) smrtcells. We sequenced one smrtcell for each Iso-seq library using sequencing kit 3.0 (#101-597-800). We then used the Iso-seq application in the Pacbio smrtlink package to generate Circular Consensus Sequences (CCSs), re-move cDNA primers and concatemers, identified strandedness, trim polyA tails, and perform de novo clustering and consensus call to output high-quality full-length consensus isoforms. Truseq stranded mRNA libraries (TruSeq Stranded mRNA LT Sample Prep Kit/TruSeq Stranded mRNA Sample Preparation Guide, Part # 15031047 Rev. E) were generated and sequenced on a Novaseq6000 S4 lane (150bp PE) at Psomagen, Inc. A total of 6 libraries were sequenced: 2 for brain, 2 for ovary and 2 for muscle RNA samples. Newly-generated IsoSeq and RNAseq data, RNAseq data from other individuals[123] (Table S4A), and protein alignments were used to guide the gene prediction process to generate the first NCBI RefSeq annotation for the species (NCBI *Hirundo rustica* Annotation Release 100) using the NCBI Eukaryotic genome annotation pipeline.[12,27] To obtain the coordinates of the different functional features of bHirRus1 (genes, exons, introns, CDS, 5′ UTR, 3′ UTR) for the following analysis, we parsed the NCBI annotation GFF3 file with GenomicFeatures[83] using a modified R script, excluding tRNAs, pseudogenes and C/V_gene_segments. Scripts used for this analysis can be found on GitHub (https://github.com/SwallowGenomics/BarnSwallow/tree/main/Analyses/GenomicFeatures).

Chromosome size estimations from karyotype images

Chromosomes sizes were estimated from four karyotype images using the chromosome_size software (https://git.mpi-cbg.de/dibrov/chromosome_size#example). The average size value was calculated for each chromosome. Sizes were correlated with the assembly chromosome sizes using Spearman nonparametric rank test.[124]

Chromosome classification assignment

We assigned bHirRus1 chromosomes to the three typical avian chromosomal groups (macrochromosomes, intermediate chromosomes, microchromosomes), adapting the classification described by the chicken genome consortium.[125] Here the authors assigned chromosomes ranging from 188 to 56.6 Mb to macrochromosomes, chromosomes from 33 Mb to 20 Mb to intermediates and chromosomes smaller than 20 Mb to microchromosomes. For the barn swallow genome, we designated chr7 (38.46 Mb) and chr8 (36.08 Mb) to the intermediate group, given their divergence in size with the larger macrochromosomes.

Assembly evaluation and comparison with other barn swallow assemblies

The commands used for the assembly evaluation can be found on the project GitHub page (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/assembly_evaluation/assembly_evaluation.txt).

Raw reads alignments

Raw PacBio subreads were converted to fastq files with samtools[78] bam2fq 1.10. Each read set was aligned to both assemblies with bwa-mem[75] 0.7.17-r1188 and then converted to bam with samtools sort 1.10 with the -o option. The coverage was calculated from the bam file with mosdepth.[84]

Assembly statistics

Assembly metrics for all the assemblies were obtained with asm_stats.sh (https://github.com/VGP/vgp-assembly/blob/master/pipeline/stats/asm_stats.sh) with the mean predicted haploid genome size from Genomescope2.0 (1,241,727,742 bp; Table S1A). Meryl[25] was used to count 21-mers from 10x linked reads that was then used in Merqury,[25] a reference-free tool that computes per-base assembly accuracy (QV), completeness and *k*-mer multiplicity. Functional completeness was evaluated with BUSCO[22,26] as already explained.

Hi-C contact heatmaps

The three-dimensional conformation of chromosomes can be visualised as Hi-C interaction heatmaps through the alignment of the read set against the assembly. Contact maps were created from bwa-mem[75] alignments with PretextMap (https://github.com/wtsi-hpag/PretextMap) and visualised with PretextView (https://github.com/wtsi-hpag/PretextView).

Masking of repetitive regions

The assemblies were soft-masked with WindowMasker 1.0.0[85] and RepeatMasker 4.1.0[86,126] (http://www.repeatmasker.org). RepeatMasker was run with NCBI/RMBLAST 2.10.0+ with Dfam_3.1 (profile HMM library) and Repbase[127] version 20,170,127 as repeat databases with the 'aves' repeat library. First, the genomes were processed separately with both tools. Then, 1-base repeat coordinates from RepeatMasker were used to further mask the Windowmasker-masked genome with bedtools maskfasta.

Chromosome size and genomic content correlations

Spearman nonparametric rank test[124] was used for the correlation between features and chromosome sizes, while Mann-Whitney U Test[128] was used to compare differences between microchromosomes and the other chromosomes. GC content was calculated with bedtools[87] nuc. CpG islands for bHirRus1 were downloaded from the UCSC browser (https://genome.ucsc.edu/cgi-bin/hgGateway). The fraction of the chromosomes covered by CG, CpG islands, genes and repeats (in percentage), was correlated with chromosome sizes (Table S2). Based on their high PacBio long-reads coverage (Table S2), microchromosomes 31, 33 and 34, representing approximately 0.2% of the assembly sequence (2.7 Mbp), were excluded from all correlation analysis.

Haplotig purging in Chelidonia

To confirm the presence of alternate haplotigs in Chelidonia and to investigate whether they affected *k*-mer and BUSCO[26,22] completeness, and increased the size of the assembly, we ran purge_dups[74] on Chelidonia with default parameters. The removal of retained haplotigs was evaluated with BUSCO,[22,26] Merqury[25] and asm_stats (https://github.com/VGP/vgp-assembly/blob/master/pipeline/stats/asm_stats.sh).

Selection analysis on multiple whole-genome alignments

Cactus alignment

Progressive Cactus[32] v1.3.0 with default parameters was used to align bHirRus1 with 10 chromosome-level annotated Passeriformes genomes available on NCBI and the Chicken genome (Table S5A). A maximum of 10 species were chosen due to the considerable computational demands of Cactus. The genomes were soft-masked with WindowMasker[85] and RepeatMasker[86] (http://www.repeatmasker.org)[32] and then aligned. Progressive Cactus[32] v1.3.0 was run with the command "cactus –logInfo –logError –binariesMode local –workDir = /data/workDir jobStore SeqFile3.txt alignment.hal". The SeqFile.txt file contained the paths to the masked assembly files of the 10 bird species (Table S5A) and the guide tree taken from TimeTree[88] (Figure S3A) in Newick format. Despite different runs with the same parameters, two species failed to align (*Parus major* and *Ficedula albicollis*) and were excluded from the subsequent analyses (Table S5A). The alignment coverage for each species was calculated with halAlignmentDepth[89] with the –noAncestors option and the barn swallow (bHirRus1) as target species. Coverage was computed for each chromosome separately and the values among different species were averaged (Table S5A). The parameter –step 200,000 was added to the command to generate track I of Figure 2C. A custom script was used to calculate the number of genomes covering each bHirRus1 chromosome base (Table S5B). More details on the commands can be found on the project GitHub page (https://github.com/SwallowGenomics/BarnSwallow/tree/main/Analyses/Cactus_alignment).

Neutral model estimation

PHAST v1.5[33] was used in combination with the HAL toolkit[89] for the selection analyses. An alignment in the MAF format was extracted for each bHirRus1 chromosome from the Cactus HAL output using hal2maf[89] with the –noAncestors and –onlyOrthologs options. The MAFs were post-processed with maf_stream merge_dups consensus (https://github.com/joelarmstrong/maf_stream), as previously described.[15] The non-conserved neutral model was trained from fourfold degenerate (4d) sites in the coding regions of the barn swallow annotation.[35,129] Briefly, CDS that fall within bHirRus1 chromosomes were extracted from the NCBI gff3 annotation file. msa_view[33] was used to extract 4d codons and 4d sites from each MAF separately, using the correspondent CDS coordinates. The combined 4d sites were used with phyloFit[33] (–subst-mod REV –EM) to generate the neutral model. The command used to estimate the neutral model can be found on GitHub (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/Selection%20analysis/neutral_model_estimation.txt).

PhyloP analysis

PhyloP[33] was run on each chromosome separately using the neutral model with LRT method and in the CONACC mode. Due to the low number of aligned species, and therefore the low total branch length between them,[15] no significant calls were found after the false discovery rate (FDR)[34] correction with 0.05 as significance level. We increased the statistical power of the constraint analysis by running phyloP on 10bp windows. Briefly, the aligned coordinates of bHirRus1 in the Cactus alignment were obtained and divided into 10bp windows. PhyloP was run again on the windows (LRT method and CONACC mode), and the FDR correction at 5% was applied. Windows smaller than 10bp were discarded and windows overlapping with assembly gaps were removed. Spearman nonparametric rank test[124] was used to correlate chromosome size and the fraction covered by phyloP sites (Table S2). Wilcoxon signed-rank test[130] was used to compare differences between microchromosomes and the other chromosomes. The commands used to perform the phyloP analysis can be found on GitHub (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/Selection%20analysis/phyloP_analysis.txt).

PhastCons analysis

An additional conservation analysis was performed using PhastCons[33] with the same neutral model as phyloP analysis, to predict discrete conserved elements (CEs). PhastCons requires parameter tuning to reach the desired levels of smoothing and coverage.[33] Given the low number of species and the high number of sites in our alignment, point 4.1 of PhastCons HOW TO guide[129] was followed. The initial length expected for phastCons was guessed at 20 bp, while the target coverage, which is the fraction of bases expected to be conserved, was set at 0.174. This value was calculated as the ratio between the expected conservation fraction (13.2%[15]) and the mean mappability between the barn swallow and the aligned genomes (76%; Table S5A). The parameters were tuned such that around 65–70% of the CDS bases were covered by phastCons conserved elements (CEs)[35,37] and the smoothing PIT was around 10.[35,129] Briefly, each chromosome MAF file extracted for phyloP analysis was split into 1 kbp chunks and 200 chunks were randomly selected from the set. PhastCons was run on each sampled chunk with the –no-post-probs and –gc 0.425 tuning options, using the initial expected length and coverage, as well as the previously generated 4d non-conserved neutral model. The parameters, initially estimated separately, were averaged with phyloBoot,[33] obtaining tuned conserved and non-conserved neutral models, which were then used by phastCons to predict conserved elements and conservation scores on each chunk. The smoothing level was checked with consEntropy[33] and coverage between CDS and the predicted CEs was manually verified. The analysis was repeated until the desired smoothing and coverage were reached (–target-coverage 0.22 –expected-length 8). Following Craig et al.,[37] windows that overlapped for more than 20% with an assembly gap were removed, and all bases that fell into gaps were filtered out. Correlations between phyloP conserved elements and phastCons CEs as the number of elements per 10kb windows were computed with the Spearman correlation rank test.[124] The commands used for this analysis can be found on GitHub (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/Selection%20analysis/phastCons_analysis.txt).

Candidate gene detection

To calculate the percentage of conserved and accelerated bases in bHiRus1 we considered how many chromosomal bases (1,082,536,200 bp) were detected as conserved and accelerated by both phyloP and PhastCons (Table S6A). To detect candidate genes, we intersected the conserved and accelerated bases detected with each annotated class extracted with GenomicFeatures. Bases overlapping with more than one feature were hierarchically assigned based on their first appearance[37,131] in this order: CDS, 5′ UTR, 3′ UTR, intronic, intergenic. Genes without identified orthologs (“LOC” genes) were discarded. The commands used for this analysis can be found on GitHub (https://github.com/SwallowGenomics/BarnSwallow/tree/main/Analyses/GenomicFeatures)

Gene ontology enrichment analysis

The gene ontology (GO) analysis was performed on the top 500 genes with the most overlaps with phyloP accelerated and conserved sites using the Generally Applicable Gene-set Enrichment (GAGE) method[90] (*gage* R package). GAGE detects enrichment for genes' functions (GO terms) in the tested datasets with respect to a broader dataset. A GO term is considered enriched in the tested dataset when the associated p value after FDR correction (q-value) is <0.05. Previous to *gage* analysis, *bioMart*[91] R package was used to retrieve correspondence between the zebra finch and human Ensembl IDs and associate the latter with GO terms. The zebra finch annotation was used as the broader complete dataset since the barn swallow could not be found on Ensembl yet at the time of the analysis. Human genes were used since annotation with GO terms should be more accurate. The script used can be found on the project GitHub page (https://github.com/SwallowGenomics/BarnSwallow/tree/main/Analyses/Gene_ontology).

camk2n2 tree construction

To look at differences in *camk2n2* transcript between species with different levels of association with humans, the transcript sequences of 38 species were downloaded from NCBI (Table S31) and aligned with Muscle on MEGA.[92] The tree was then generated using the Maximum likelihood method, a generalised time reversible (GTR) model and a gamma distribution (G) with 5 categories (see Data S1).

SNP catalog generation

Datasets used

To generate the catalog of genetic variants, five Italian barn swallow individuals were sampled. HMW DNA was extracted from the blood samples and sequenced with PacBio HiFi technology (see "HiFi reads processing for SNP catalogue, titration and phasing experiment" section for a detailed description of the generation and processing of HiFi data). Then, all publicly available datasets (Table S12) were used to complement our newly generated HiFi reads set and generate a comprehensive genetic marker catalog for the barn swallow. Raw reads from public datasets were downloaded using fasterq-dump v2.9.1 from SRA Toolkit (https://github.com/ncbi/sra-tools). The data were single-end, except WGS data in ds2 and ds3.1 and ddRAD data in ds5. Quality control was performed on all raw reads using Fastqc v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Multiqc v1.9[93] (https://github.com/ewels/MultiQC). Low quality bases were trimmed using Cutadapt v2.10[94] (Figure S8). BBDuk, from BBMap v38.18[95] was used to remove Illumina adapters (k = 23, max mismatches = 1). Fastq files were aligned to bHirRus1 reference genome using bowtie2 v2.4.1.[68] The unmasked genome was used as reference. For WGS data, duplicated reads were removed using the Picard MarkDuplicates tool v2.23.4 (http://broadinstitute.github.io/picard/). Samtools v1.9[79] (https://github.com/samtools/samtools) was used to sort and index alignments. Alignment files generated from paired-end genomic data were further processed with Bam clipOverlap software v1.0.14 (https://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap) to trim overlaps between paired reads. The complete pipeline used to download and align reads is available on the project github page (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/popgen_data_download_alignment/popgen_pipeline.bash).

Variant calling and filtering

Freebayes v1.3.1[77] (https://github.com/freebayes/freebayes) was used to call variants. To reduce computational time, a script adapted from the VGP assembly pipeline (https://github.com/VGP/vgp-assembly/blob/master/pipeline/freebayes-polish/freebayes_v1.3.sh) was used to parallelize the process by subsetting the reference genome by scaffolds. Variants were called with the options –min-mapping-quality 10 –min-base-quality 20 –populations (all other parameters were left to default). Due to the lower sequencing coverage, –min-alternate-count 0 was used for ds6. The coordinates of the repetitive regions were extracted from the masked reference genome with a python script (https://gist.github.com/danielecook/cfaa5c359d99bcad3200) and the unmasked regions identified with bedtools v2.29.2[87] using the complement command. All vcf files were first filtered to remove variants falling within repetitive regions, multiallelic SNPs and indels. Variants were then split by population, and further filtering steps and thresholds are detailed in Table S21. We removed sites showing more than twice the mean read depth across samples (INFO/DP field). In the vcf generated by Freebayes, genotype quality is expressed as QR (quality reference) and QA (quality alternate). We marked as missing all genotypes in which both values were below the threshold reported in Table S21. For FMT/DP filtering, we used as maximum value twice the average DP value and we approximated the 5% quantile of the distribution to set the minimum value. Individuals presenting a high amount of missing data (>70%) were discarded (Table S21). Variants were also filtered for minor allele frequency (maf) with the usual 5% threshold and average fraction of missing sites among individuals (Table S21). All filters were applied using bcftools v1.1[79] (https://github.com/samtools/bcftools) with the view and filter commands, except the removal of variants falling within repetitive regions, performed with bedtools v2.29.2[87] using the intersect command and the coordinates of the unmasked regions previously identified. Standard statistics from the vcf files (in particular average site depth and average individual depth) were calculated using VCFtools v.0.1.16[96] (https://github.com/vcftools/vcftools). An example of the complete set of commands used to filter variants (from ds2.2) can be found here (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/variants%20filtering/filtering_commands.txt).

To compare variant identification achieved with a linear genome (bHirRus1) and with the pangenome, we used the raw vcf file generated by Freebayes with the options –min-mapping-quality 10 –min-base-quality 20, extracting the 16 Illumina WGS samples relative to ds3.1. Only biallelic SNPs were kept for the comparison. Bcftools v1.1[79] (https://github.com/samtools/bcftools) was used to manipulate the vcf file and extract the genomic region corresponding to the *camk2n2* gene. To validate variants from reads aligned to bHirRus1, IGV[97] was used for visual inspection.

SNP statistics and correlations with genomic features

For all the analyses described in this subsection and the following one ("SNP density plotting"), all datasets generated with the same sequencing technology were combined (HiFi WGS; Illumina WGS; Illumina ddRAD). SNP density for each chromosome (excluding unlocalized/unplaced scaffolds) was computed on 10 kbp windows and SNPs were counted using bedtools v2.29.2[87] with the coverage -counts option. The average SNP density values across all chromosomes for each sequencing technology was calculated in R using the weighted mean function. Mean value was weighted for the window size to take into account truncated windows potentially present at chromosome ends. For the HiFi dataset (ds1) also a 5x downsampled HiFi dataset was generated (see "HiFi reads processing for SNP catalogue, titration and phasing experiment" section, "titration of HiFi reads" subsection, first titration experiment) considering the 20x read coverage of each sample (except for the A2 sample, starting from 15x) as the truth set (variants from the 5x reads set were intersected with variants from the 20x reads set using bedtools v2.29.2[87] with the intersect command).

For each chromosome and dataset, SNPs falling in intervals corresponding to genic, intergenic, exonic and intronic regions as determined from NCBI annotation were counted using bedtools v2.29.2[87] with the coverage -counts option (Data S1). To analyze correlations between SNP density and GC content in our catalog, the GC content was calculated using bedtools v2.29.2[87] with the -nuc option on 10 kbp windows and SNPs were counted every 10 kbp window. Correlation was tested in R computing the Spearman nonparametric rank test[124] with the R function cor.test. Unlocalized/unplaced scaffolds were excluded from the analysis. bedtools v2.29.2[87] was used to divide the genome in 10 kbp windows, using the makewindows command with the -w 10,000 flag.

SNP density plotting

To plot SNP distribution across chromosomes, SNP density was computed over 40 kbp intervals with the R[117] package *karyoploteR.*[98] Additional tracks included repetitive regions, GC content, raw reference reads coverage and assembly gaps. Repeats were annotated by Windowmasker 1.0.0[85] and Repeatmasker 4.1.0.[86,126] GC content was calculated using bedtools v2.29.2[87] with the -nuc option on 1 kbp windows. Per base coverage of raw reference reads was calculated by aligning reads back to the bHirRus1 assembly and using bedtools v2.29.2[87] with the genomecov -d option. Values were then averaged every 500 bp (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/coverage_analysis/avg_coverage.bash). Standardised values were attributed to specific coverage intervals: 0 for low coverage (between 0 and 10), 100 for regions showing twice the average coverage value (95), or higher, and intermediate fixed values for coverage between 10 and 95. Assembly gaps were removed from computation of GC content, repeat content and PacBio reads coverage.

Linkage disequilibrium and haplotype statistics analysis

Genome-wide LD decay

LD decay was evaluated in all Illumina WGS datasets using r^2 from Freebayes v1.3.1[77] variant calls. r^2 values were calculated using Plink v1.9.[99] To estimate LD decay trend across the whole genome in filtered ds2 and ds3.1, we considered marker pairs within a 55 kbp distance with the option –bcf file.bcf –r2 dprime yes-really –ld-window 999,999 –ld-window-kb 55 –ld-window-r2 0 –allow-extra-chr –out LD55kb. Option –ld-window 999,999 is required to consider variant pairs more than 9 lines apart from each other.[132] To calculate average r^2, SNP pairs were grouped according to their distance in bins of 1 kbp (range 1–55 kbp) using a custom perl script (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/LD-scripts/LDaverage.pl), that was run on Plink output. The same approach was used to calculate average r^2 values per chromosome group (macrochromosomes, intermediate and microchromosomes), except that values were then averaged across specific distance bins. Sex chromosomes were excluded from the chromosome group LD analysis.

Relationship between LD and distance from chromosome ends

A potential correlation between LD and distance from chromosome ends was evaluated in ds2.1, 2.2, 3.1.1, 3.1.2 combining chromosomes together according to their type (macrochromosomes, intermediate and microchromosomes; Data S1). Plink v1.9[99] was used to estimate r^2 values from each dataset with the option –bcf file.bcf –r2 dprime yes-really –ld-window 10,000 –ld-window-kb 20 –ld-window-r2 0 –allow-extra-chr –out LD_20kb. Then, to calculate average LD values for every marker pair having a certain distance bin from chromosome end, a custom perl script was used (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/LD-scripts). Marker pairs were grouped using 10kb as distance bin value from chromosome ends. The correlation between distance and LD values was tested in R computing the Spearman nonparametric rank test[124] with the R function cor.test.

LD scans

Before performing the LD scans, variants were filtered with bedtools v2.29.2[87] using as maximum coverage (95x) twice the average PacBio reads coverage genome wide (47.7x) and 10x as the minimum, so to ensure the exclusion of SNPs falling within collapsed or ambiguous regions of the genome. For the first LD scan, we ran Plink v1.9[99] on Illumina WGS data from American and Egyptian samples (ds3.1) considering marker pairs within a 15 kbp distance maximum, with the options –bcf file.bcf –r2 dprime yes-really –ld-window 10,000 –ld-window-kb 15 –ld-window-r2 0 –allow-extra-chr –out LD15kb. To scan for genes showing high LD values, r^2 was chosen as it is generally more informative for small datasets and also more consistent with allele frequency variation,[133] whereas D' can be more prone to inflation. To compute the average LD, each scaffold was divided in sliding non-overlapping 5 kbp windows with a custom perl script (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/LD-scripts/chr_ld.pl), requiring a minimum of 100 markers per window. Only genomic windows with average $r^2 > 0.3$ were extracted (Table S22). The threshold was chosen based on similar studies.[133,134] Coordinates were intersected with the NCBI annotation to find genes potentially carrying alleles with high LD using bedtools v2.29.2.[87] For further analysis, two 5 kbp intervals were joined into the same ROI if the distance between them was lower than 100 kbp. Intervals showing high LD values were excluded if in proximity (within ~5 kbp) of potentially collapsed or low-confidence assembly regions (considering a PacBio reads coverage value higher than twice the average genome-wide coverage or lower than 10, respectively) or if not carrying any annotated gene. For the average LD computation of chr6 in the *H. r. savignii* (ds3.1.1) and *H. r. erythrogaster* (ds3.1.2) populations separately we used the procedure described above but requiring a minimum of 10 markers per window. The *bdnf* gene region (belonging to ROI 45) was then analyzed in more details, and LD heatmaps were generated using LDBlockShow v1.36[100] (https://github.com/BGI-shenzhen/LDBlockShow) with the options -InVCF file.vcf -OutPut Scaffold_name -Region Scaffold:start-end -OutPng -SeleVar 2. CpG islands along the *bdnf* sequence were identified with cpgiscan v1.0[101] (https://github.com/jzuoyi/cpgiscan), combining neighboring CpG islands when their distance was lower than 100 bp (Data S1 and Figure S9).

iHS computation

To calculate iHS, namely the standardised log-ratio of the iHH (integrated haplotype homozygosity) values for the two alleles, variants present on chr6 were phased with WhatsHap v0.18[102] (https://github.com/whatshap/whatshap) and the Rehh[103] R package was used (Data S1). Before iHS computation, variants were filtered to remove sites showing a fraction of missing genotypes across samples higher than 0.1 and sites with maf <5%, using Rehh filtering options min_perc_geno.mrk = 90 and min_maf = 0.05. Extended haplotype statistics were then calculated using the scan_hh (with the polarised = FALSE option) and the ihh2ihs (setting freqbin = 1) functions. To perform FDR correction, the qvalue R package was used (https://github.com/StoreyLab/qvalue). This analysis was performed on populations relative to ds3.1, ds2.1 and ds2.2. The complete list of commands used for iHS computation can be found here (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/iHS%20analysis/iHS_analysis_script.R).

HiFi reads processing for SNP catalog, titration, and phasing experiment

HiFi reads alignment, variant calling, and filtering

HiFi reads from ds1 samples were aligned to bHirRus1 with pbmm2 v1.3.0 (https://github.com/PacificBiosciences/pbmm2) using default parameters for PacBio CCS reads with the options align –preset CCS –sort -j 32 –log-level INFO reference.mmi reads.ccs.bam file.aligned.ccs.bam. The genome-wide coverage of mapped reads was computed with bedtools v2.29.2[87] using the genomecov command. At first, alignments were used to call small variants using DeepVariant v1.0.0[104] (https://github.com/google/deepvariant) with default parameters for PacBio reads individually for each sample. Variants were first filtered to remove multiallelic SNPs and indels. SNPs falling within repetitive regions were removed as described for the publicly available datasets. Next, only SNPs with a genotype quality value higher than 20 were kept, and 5% and 95% quantiles of the read depth values distribution were used to set the minimum and maximum site coverage. Filters were applied using bcftools v1.1,[79] and filtered variants from each sample were merged with the same tool to estimate and plot SNP density across chromosomes as described for Illumina WGS and ddRAD data. These HiFi variants were included in the genetic marker catalog (Figure 3B). For the comparison between Illumina and HiFi technology, Samtools v1.9[79] was used with the view command and the -q flag to exclude reads with a mapping quality value lower than 30 (for Illumina data) and 60 (for HiFi data), based on Hon et al.[135] The proportion of the genome covered by the alignment was computed with bedtools v2.29.2[87] with the genomecov -bg option. All bases with read depth ≥ 1 were extracted from bedtools output. HiFi joint variant calling of SNVs and indels was performed using gVCF files from DeepVariant v1.1.0[104] per-sample calls, jointly called with GLNexus[105] pipeline (https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake). For joint calling of SNVs and indels, DeepVariant v1.1.0[104] was run twice, the second time after an intermediate variants phasing step performed with WhatsHap v1.0.[102] For SVs, pbsv v2.6.0[106] (commit v2.4.1–155-g281bd17) (https://github.com/PacificBiosciences/pbsv) was used for per-sample and joint variant calling. The minimum SV length was set to 20 bp.

The raw variant calls obtained with DeepVariant from ds1 were also used to confirm the SNPs identified within the pangenome. Only biallelic SNPs were kept for the comparison. Bcftools v1.1[79] (https://github.com/samtools/bcftools) was used to manipulate the vcf file and extract the genomic region corresponding to the *camk2n2* gene.

Titration of HiFi reads

Two downsampling experiments were conducted (Data S1), the first one after individual variant calling and the second one after joint variant calling (N = 5). For the individual titration experiment, all HiFi reads were first downsampled to 20x coverage using Rasusa v0.3.0[107] (https://github.com/mbhall88/rasusa), except for the A2 sample where the sequencing coverage was 15x. Three different truth sets were generated, first (truth set 1) using the vcf file derived from the 20x coverage alignment of each sample; second (truth set 2) by intersecting this 20x file with a set of publicly available barn swallow variants (dst3.1); third (truth set 3) from the intersection of all variants from the 5 samples at full sequencing coverage. Each read set was further downsampled at 15x, 10x and 5x, in triplicate for each condition. Reads were aligned to bHirRus1 and variants were called as described in the previous subsection for per-sample variant calling. Specific filters were applied as described in the previous subsection. The three different truth sets were then intersected with the variants recovered after every titration using bcftools v1.1[79] with the isec command and the -w1 flag. Recall rate, precision and F1 score were estimated for each titration experiment. The recall rate at the different coverage values was estimated as the number of shared variants after intersection divided by the total number of variants in the truth set for each sample, while the precision rate was estimated as the number of shared variants after intersection divided by the total number of variants identified in each particular titration replicate. The F1 score, the harmonic mean between recall rate and precision rate, was estimated as $F1 = 2 \times \frac{precision \times recall}{precision + recall}$. For the second titration experiment, reads were randomly downsampled using Rasusa v0.3.0[107] tool as described above for the first experiment. Reads were then aligned to bHirRus1 using pbmm2 v1.4.0, variants were called as described in the previous subsection for joint variant calling and recall rate was estimated considering the full-coverage joint calling as truth set.

Phasing of HiFi read sets

Variants obtained with HiFi reads (ds1) were filtered to remove multiallelic SNPs and indels. Only SNPs with a genotype quality value higher than 20 were kept, and 5% and 95% quantiles of the read depth values distribution were used to set the minimum and maximum site coverage. Next, to estimate and plot haplotype-phased blocks length across chromosomes, variants were phased using WhatsHap development version v.1.2.dev2+g3dffe4a[102] with the options stats –chr_lengths –tsv (Data S1).

Pangenomics

Generation of the pangenome

For the generation of the pangenome, we used our newly generated HiFi data from the five *H. r. rustica* barn swallow individuals (ds1). HiFi reads were checked for adapter contamination and trimmed accordingly with cutadapt v3.2.[94] Genomescope2.0[21] was used to predict assembly statistics from HiFi raw data (Table S11D). Hifiasm v0.13-r307[52] was used to assemble both primary and alternate assemblies which were then purged using purge_dups[74] with the minimap2 option -xasm20 and custom cutoffs (Table S11E).[136] The two cutoffs were calculated starting from the *k*-mer coverage (kcov) computed by Genomescope2.0[21] (value1 = kcov*1.5, value2 = value1*3). The assemblies were masked with WindowMasker 1.0.0[85] and RepeatMasker 4.1.0[86] to reduce the alignment computational time.[32] The Cactus Pangenome Pipeline included in Cactus[32] v1.3.0 was run as described in the software documentation (https://github.com/ComparativeGenomicsToolkit/cactus/blob/master/doc/pangenome.md). Briefly, Minigraph[108] v0.14-r415 was used to generate a GFA graph starting from the purged HiFi primary and alternate assemblies (Table S11F) and bHirRus1 primary and alternate assemblies with the -xggs preset. Then, cactus-graphmap was used to align the input fasta sequences to the minigraph.Cactus-align was then used to run Cactus in pangenome mode to generate both a HAL alignment and a vg graph starting from the previous alignment. The vg file was modified using vg mod -O for a better visualisation of paths. The commands used for the assembly of the pangenome and subsequent ortholog analysis can be found on the project GitHub page (https://github.com/SwallowGenomics/BarnSwallow/tree/main/Analyses/Pangenome).

Pangenome ortholog analysis

Orthologous genes were found running HALPER[109] following the steps described on GitHub (https://github.com/pfenninglab/halLiftover-postprocessing). Briefly, from the HAL alignment, the coverage of bHirRus1 was calculated with halAlignmentDepth.[89] Then, a file for the ortholog extension was generated from the coverage file and halLiftover[89] and used to lift bHirRus1 gene coordinates on the alternate assembly and the HiFi assemblies aligned in the pangenome graph. Orthologs were then found using the lifted genes. The resulting lists of orthologs were manually evaluated to find genes shared between individuals. The 234 genes that were found only in the bHirRus1 assembly were searched in the HiFi raw reads with BLAST 2.10.1+.[81] The alignments were checked to find genes present for more than 80% of their sequence in the reads and 99% identity with the query sequence. To assess whether the missing genes in bHirRus1 after the raw reads analysis (155) were real gene losses or related to sequencing biases in PacBio sequencing, the GC content was calculated using custom scripts and GA, GC and AT dinucleotides presence was measured as described in,[137] using sliding 128 bp windows. The Mann-Whitney U Test[128] was used to detect an enrichment in GC content in the 155 genes with respect to the other bHirRus1 genes, whilst a Chi-squared test[138] was used to detect an enrichment in CG, GA and AT dinucleotides. To account for GA presence on both strands, GA and TC dinucleotides were added together.

Comparison between variants embedded in the pangenome and variants called with deepvariant

The SNPs found between the haplotypes included in the pangenome were manually detected looking at the graphical representation of the pangenome in *camk2n2* region (Figure 5F). SNPs called with deepvariant using the HiFi reads and the linear reference genome (see section 'HiFi reads processing for genetic variants identification') in *camk2n2* regions were retrieved from the whole VCF before filtering (no filtering was performed for the pangenome variants). Only SNPs were retained, excluding indels and reference calls (Table S19).

Pangenome variant calling

The pooled Illumina WGS data for 16 barn swallow individuals[2] (ds3.1) were aligned against the pangenome graph using vg map,[59] after some steps of pre-processing with vg mod -X 256 and vg prune -k 45. The samples were not separated (~5x) to simulate the alignment of an individual with high coverage. The subgraph representing *camk2n2* coordinates was extracted with vg chunk (pg, packed-graph format) and the aligned reads (gam format) were embedded in the subgraph using vg augment, generating augmented pg and gam files. Snarls were computed separately with vg snarls from the augmented vg, while the read support was computed from the augmented gam with vg pack. Variants were called with vg call. The commands used can be found on GitHub (https://github.com/SwallowGenomics/BarnSwallow/blob/main/Analyses/Pangenome/Pangenome_variant_calling/Variant_calling.txt). Variants were filtered removing indels, 'lowad' and 'lowdepth' variants and compared to variants called with the linear reference genome. In addition, SNPs called as heterozygous with only one read supporting the alternate allele were not considered, for a more informative comparison with the variants set obtained with Freebayes using bHirRus1 as reference (where this parameter was left to the default value of 2).

Graphical representations

The R[117] package *ggplot2*[110] was used to generate correlation plots (Figures 2B and S2), histograms (Figures 5B and S3B–S3D) and the gene presence-absence matrix (Figure 5B). The R package *circlize*[111] was used to generate Circos plots and the figure legend was generated using the *ComplexHeatmap*[112] package (Figures 2C, 5A and S1). SequenceTubeMap[113] was used to graphically represent pangenome regions (Figures 5F and S7). MEGA X software[92] was used to generate the phylogenetic trees (Figure 3A and STAR Methods). The Hi-C contact heatmaps were visualised with PretextView (https://github.com/wtsi-hpag/PretextView, Figures 1D–1F). The *k*-mer profiles were generated with Genomescope2.0[21] (http://qb.cshl.edu/genomescope/genomescope2.0/) and Merqury[25] (Figures 1B and 1C). Snail plots were generated with BloobToolKit[114] (Figure 1G). Alignment dot plot was generated with D-genies[115]

(Figure 1H). Manhattan plots were generated with the R package *CMplot*[116] (Figures 3E and 3F). IGV[97] was used to visualise aligned features to the genome (Figure S4). R[117] package *karyoploteR*[98] was used to plot SNP density visualisation across all chromosomes (Figures 3B, S5 and S6). SNP density was computed using the internal function kpPlotDensity using 40 kbp as window size, for the three types of sequencing technologies considered. To plot SNPs distribution across all chromosomes for the 5x downsampled HiFi dataset (Figure S6), the 20x read coverage of each sample (except for the A2 sample, starting from 15x) was used as the truth set (variants from the 5x reads set were intersected with variants from the 20x reads set before plotting). Both coverage and GC content were plotted with the kpHeatmap function. The heatmap relative to Pacbio coverage was generated using the *viridis* package. Repeats and assembly gaps were plotted using the kpPlotRegions function. Only repeats larger than 3 kbp (larger than 1 kbp for Figure S5, relative to microchromosomes) were plotted. The figure legend was generated using the *ComplexHeatmap*[112] package. Unlocalized/unplaced scaffolds were excluded. The R package *ggplot2* was used to plot genome-wide LD decay (geom_line function) and LD per chromosome group (geom_boxplot function) (Figure 4). After LD scans, LD values were plotted with the KaryoploteR[98] package using the kpPoints and kpLines functions. SNP counts for the two populations were plotted with the kpHeatmap function. The *bdnf* transcript isoforms structure was drawn using the *ggplot2* package. IGV[97] was used to visualise *bdnf* region containing previously annotated methylation sites from the Cactus multialignment (Figure S9C).

The map showing sampling locations from all datasets was generated in R using the packages *ggplot2,*[110] *rnaturalearth*, *sf* and *rnaturalearthdata* (Figure S3A). Average LD values at increasing distance from chromosome ends were plotted with the *ggplot2*[110] package using the geom_point function and combined together with the ggarrange function (Figure S11). iHS values were plotted using the manhattanplot function of the *Rehh*[103] package (Figure S12). Histograms of the HiFi reads coverage were generated with the *ggplot2*[110] package using the geom_bar function (Figures S13A–S13E). To plot recall rate values after HiFi titration experiments, the functions geom_line and geom_point of the *ggplot2* package were used. For the second titration experiment, the legend was generated using the *ComplexHeatmap*[112] package and plots were arranged together with the packages *grid* and *gridExtra* (Figures S13G–S13I). Before plotting phased blocks length, the WhatsHap development version v.1.2.dev2+g3dffe4a[102] command stats –gtf was used to generate a.gtf file with the size and position of the phased blocks. Phased blocks computed from HiFi reads were plotted with the *KaryoploteR*[98] package using the kpRegions function (Figure S14A). The percentage of phased chromosomes, colored by type, averaged across samples, was plotted with the *ggplot2*[110] function geom_boxplot (Figure S14B). See this github section (https://github.com/SwallowGenomics/BarnSwallow/tree/main/Plots%20and%20figures) to retrieve the lists of commands used for all figures and plots.

Cell Reports, Volume *42*

Supplemental information

A chromosome-level reference genome and pangenome for barn swallow population genomics

Simona Secomandi, Guido R. Gallo, Marcella Sozzoni, Alessio Iannucci, Elena Galati, Linelle Abueg, Jennifer Balacco, Manuela Caprioli, William Chow, Claudio Ciofi, Joanna Collins, Olivier Fedrigo, Luca Ferretti, Arkarachai Fungtammasan, Bettina Haase, Kerstin Howe, Woori Kwak, Gianluca Lombardo, Patrick Masterson, Graziella Messina, Anders P. Møller, Jacquelyn Mountcastle, Timothy A. Mousseau, Joan Ferrer Obiol, Anna Olivieri, Arang Rhie, Diego Rubolini, Marielle Saclier, Roscoe Stanyon, David Stucki, Françoise Thibaud-Nissen, James Torrance, Antonio Torroni, Kristina Weber, Roberto Ambrosini, Andrea Bonisoli-Alquati, Erich D. Jarvis, Luca Gianfranceschi, and Giulio Formenti

Supplementary Information

Data S1 – Supplementary results

1 - bHirRus1 manual curation

1 - 1 First curation step

A first round of manual curation performed at the end of the assembly pipeline (Figure 1A) introduced 174 rearrangements by breaking and joining scaffolds and resulted in the removal of 8 false duplications. This reduced the genome length by 3.2 Mb, decreased the scaffold count by 14% to 578 and increased the scaffold N50 to 76.2 Mb (+23%; Table S1a). Overall, 98.27% assembled sequences could be assigned to chromosomes (Table S2).

1 - 2 BUSCO genes recovery

The comparison between curated bHirRus1 and Chelidonia[1] revealed a discrepancy in BUSCO[2–4] gene completeness. Chelidonia overall completeness was 95.9%, while that of the curated bHirRus1 was 94.8%, with 139 missing genes (Table S1d). A slight decrease in BUSCO scores in highly contiguous assemblies generated with Hi-C scaffolding has been previously observed (e.g. in the yellow perch[5]). The difference is not necessarily due to a higher gene fragmentation, but to BUSCO mode of operation in relation to the identification of the best candidate genomic regions across different scaffolds. In a second step of manual curation, BUSCO results were carefully analysed in order to improve the functional completeness of our assembly. In bHirRus1, we found 13 of the BUSCO genes missing in Chelidonia (Table S23). However, of the 139 missing BUSCO genes in bHirRus1 (Table S1d), 53 were identified in Chelidonia, 21 in the alternate assembly (14 in common with those in Chelidonia) and one in the VGP assembly pipeline intermediate p1 (Tables S24 and S25). According to the BLAST search of the gene sequences extracted from Chelidonia, the alternate assembly, and p1, 19 of those genes were indeed present in bHirRus1 (Tables S24 and S26a,b,c), implying that BUSCO results were partially inaccurate, and several genes that should have been identified as present were missed possibly because of automated gene prediction issues. In addition, 7 out of the 9 missing alternate genes in bHirRus1 (g20-25, g27; Table S24) were found by BUSCO in the primary contigs (c1) at the beginning of the VGP assembly pipeline (Figure 1A), and in the alternate haplotig (p2) after haplotig purging with purge_dups[6], but not in the primary purged contigs (p1), showing that while purge_dups removes false duplications and repeats, it can also remove some coding genomic regions. In particular, 6 of the removed regions containing the genes were flagged as haplotigs and one as repetitive by purge_dups (Table S26d). A manual evaluation of the alignments between bHirRus1 and Chelidonia, alternate assembly and p1 contigs or scaffolds containing the missing genes, further confirmed the presence of the genes and the absence of the missing ones in our VGP assembly (Table S27). One gene (g1; Table S27a) extracted from Chelidonia partially aligned to bHirRus1 and therefore was not recovered to avoid introducing duplications. The sequences of the missing genes were extracted from the alternate assembly and from Chelidonia together with the entire scaffold fragment that did not contain significant alignments to bHirRus1. Thirty-two sequences containing 35 of the missing gene sequences were extracted from Chelidonia, and 7 sequences containing one gene each

were extracted from the alternate assembly (Table S28). After a careful evaluation, two of the 9 alternate assembly genes that were missing from bHirRus1 (g27-28; Table S25b) were recovered from Chelidonia instead, and some scaffolds were trimmed to avoid introducing duplicated sequences (See Notes column in Table S28). The recovered sequences were renamed accordingly and added to bHirRus1 to generate a more complete assembly. In total, 42 genes were recovered. The BLAST search was repeated on the updated assembly and confirmed the presence of the recovered genes (Table S29). BUSCO completeness of the updated assembly was 96%, slightly higher than that of Chelidonia (95.9%; Table S1d). After the gene recovery, the missing genes decreased from 139 to 97 in bHirRus1 (-1.3%; Table S1d). *K*-mer completeness increased from 33.2% to 33.3% in the new version with respect to the original, while QV slightly decreased from 44% to 43.7%, due to the introduction of relatively lower quality sequences in the recovered genes (Table S1c). Duplicated genes minimally increased as a result of the process (from 0.46% to 0.49%; Table S1c).

2 - Barn swallow mitogenome

The mitogenome sequence included in the VGP assembly pipeline for polishing, is 16,277 bp long (Figure S1). Our annotation included 2 rRNA genes, 20 tRNA genes and other 12 genes, in line with closely related birds[7] (Table S30). We could not resolve the CR region due to the presence of repetitive elements, but this is presented for another barn swallow individual in our companion paper[8].

3 - Karyotype reconstruction

The karyotype of the barn swallow is composed of 2n = 80 chromosomes, which corroborate previous studies[9,10] (Figure 2A). The karyotype is made of 7 pairs of macrochromosomes (pairs 1-6 and the Z chromosome), 7 pairs of chromosomes of intermediate size (pairs 7-13) and 26 pairs of microchromosomes (pairs 14-39). The morphology of the macrochromosomes is similar to that of other swallow species[11]. In particular, pairs 1, 4, 10 and the Z chromosomes are metacentric, pairs 2, 3 and 5 are acrocentric, while pair 6 is submetacentric. Microchromosomes morphology is not identifiable due to their small size.

4 - bHirRus1 evaluation

Starting from unassembled 10x Linked-Reads, Genomescope v2.0[12] predicted a genome size of 1.24 Gbp, a repeat content of 281 Mbp and a heterozygosity of 1.04% (Figure 1B, Table S1a). The final assembly size is 1.11 Gbp (Table S1b), slightly smaller than expected (1.24 Gbp; Figure 1B, Table S1a). However, likely due to the high heterozygosity and possibly due the bias often introduced during 10x library preparation, the model fit from Genomescope v2.0 was only 88.8%, and therefore the estimations may be inaccurate (Figure 1B, Table S1a). The Hi-C contact heatmap confirmed the chromosomal structure of the assembly (2n = 78 + ZW), showing a strong correlation between read pairs and the corresponding scaffold sequence (Figure 1D). The final scaffold N50 and NG50 are 76 Mbp and 73 Mbp, respectively, with a maximum scaffold length of 156 Mbp (Table S1b, Figure 1G). The primary assembly also includes 1,719 contigs with a NG50 of 2.8 Mbp and a NG50 of 2.3 Mbp, 1,103 gaps with a max gap length of 859 Mbp and a N50 of 67 Mbp (Table

S1b). According to *k*-mer-based metrics[13], bHirRus1 has a per-base consensus accuracy of Q43.7 and a *k*-mer completeness of 83.3%, which reaches 93.9% when combined to the alternate assembly (74.2%) (Table S1c). Duplications in the original assembly after FALCON+unzip (c1) were 4.0% and decreased to 0.50% after purging with purge_dups[6] (p1) during the assembly pipeline, and to 0.49% in the final assembly after curation (Figure 1C and S10, Table S1c). BUSCO[2–4] completeness is 96% in bHirRus1, comprising 95.2% of complete single-copy *Vertebrata* orthologous genes and a 0.8% of duplicated complete genes, while 1.1% of identified BUSCO genes are fragmented and 2.9% are missing (Figure 1G; Table S1d). Repeat masking of the new assembly revealed 271 Mbp of repeats (Table S1), close to the Genomescope v2.0 prediction of 280 Mbp (Table S1a).

5 - Comparison between bHirRus1 and Chelidonia

Compared to Chelidonia[1], bHirRus1 is 108 Mbp smaller (1.11 Gbp vs. 1.21 Gbp) and slightly more fragmented, containing a larger number of scaffolds (617 vs. 364) and contigs (1,677 vs. 1,355; Figure 1G, Table S1b). Since the same data was used to generate the contigs, this is likely due to the different assembler employed (Canu[14] in Chelidonia, FALCON-Unzip[15] in bHirRus1). However, the scaffold contiguity is significantly higher in bHirRus1 (scaffold N50 76 Mbp vs. 26 Mbp, scaffold NG50 73 Mbp vs. 26 Mbp; Figure 1G, Table S1b). While the scaffold NG50 increased in bHirRus1, the contig NG50 reduced from 5.6 Mbp in Chelidonia to 2.7 Mbp, in line with the increase in gaps number (1,103 vs 971) (Table S1b). However, the total gap size was lower in bHirRus1 than Chelidonia (25.5 Mbp vs 40 Mbp, Table S1b). Merqury[13] computed a QV of 34 for Chelidonia, an order of magnitude less than that of bHirRus1 (Table S1c). Conversely, *k*-mer completeness in bHirRus1 is lower than that of Chelidonia (83.2% vs. 84.9%), but only when the alternate assembly is left out (Table S1c). In this case, the higher *k*-mer completeness is likely due to the presence of false duplications, i.e. duplicated copies of highly divergent heterozygous regions. Indeed, the haplotig purging in Chelidonia only consisted in discarding small contigs identified as haplotigs by BLAST searches[1]. Merqury spectra-cn plot generated for Chelidonia, revealed that the duplication content of Chelidonia is three times higher than that of bHirRus1 (1.3% vs. 0.49%; Figures 1C and S10E, Table S1c), and contains more duplicated BUSCO genes (1.2% vs 0.8%; Figure 1G, Table S1d). The haplotig purging performed on Chelidonia, reduced its size by 55 Mbp (1.16 Gbp vs. 1.21 Gbp), leading to an assembly size closer to that of bHirRus1 (1.11 Gbp; Table S1b). Twenty-seven Mbp were flagged by purge_dups as haplotigs, 16 Mbp as repeats, 10 Mbp as overlaps and 1.7 Mbp as high-coverage regions (Table S1e). These results led us to conclude that the larger assembly size and completeness of Chelidonia are largely explained by retained alternate haplotigs. Haplotigs removal was confirmed by the reduction of the duplication content in the purged assembly (from 1.3% to 0.55%; Figure S10E and S10F, Table S1c), which remains higher than that of bHirRus1 (0.49%). BUSCO identified 13 duplicated genes less (28 vs. 42), while minimally affecting the overall completeness (95.8% vs. 95.9%; Table S1d). The *k*-mer completeness of Chelidonia after purging decreased from 84.9% to 84.2%, whilst the QV increased from 34 to 34.2 (Table S1c). Moreover, Chelidonia has a higher repeat content (315 Mbp) than bHirRus1 (271 Mbp), according to Windowmasker[16] and RepeatMasker[17], also higher than Genomescope2.0[12] prediction (280 Mbp; Figure 1G).

6 - PhyloP analysis and candidate accelerated and conserved genes

The rate of accelerated and conserved sites is lower than that detected in other bird species[18–20], most likely because of the lower number of species included and the shorter total branch length between the aligned species[18]. Applying a more conservative Bonferroni correction[21] on the FDR corrected bases, we detected ~64 kbp (~0.01%) of significantly accelerated sites and no significantly conserved ones (Table S6a). Bonferroni-corrected accelerated bases, being a subset of FDR-corrected ones, follow the same pattern of overlaps with genomic features (Table S6a). The 20 genes with the most overlaps with phyloP accelerated sites (Table S7) comprise genes encoding for mitochondrial (*mrpl55*[22], *coq10a*[23]) and ribosomal proteins (*rps23*[24]), genes involved in hair growth and morphogenesis (*hoxc12*[25], *hoxc11*[26], *snrpe*[27]), neuron survival and migration (*znhit3*[28]), oxidative stress response (*bfr2*[29]), inflammatory response (il23a[30–32]), neural development (*rbm8a*[33]), cognitive deficits (*znf653*[34]), photoreception (*gngt2*[35]), angiogenesis and vascular remodelling (*acvrl1*[36–39]), embryonic stem cell differentiation (*polr3gl*[40]) and stem cell self-renewal (*mindy1*[41]), metabolic processes (*scly*[42,43]), ion transport (*kcne3*[44,45]), and pain transmission and emotional processing (*kcnip3*[46]). The 20 genes with the most overlaps with phyloP conserved sites (Table S9) are largely involved in neural development and differentiation (*nfia*[47], *sox2*[48], *cnot2*[49], *ube2d2*[50], *ube2d3*[50], *gjd2*[51], *foxp2*[52], *amd1*[53], *fgf12*[54], *hmgn3*[55], *pou2f1*[56], *ube2n*[50], *lmo4*[57], *mosmo*[58]). The top candidate gene with the most overlaps between its CDS and phyloP conserved sites is *camk2n2* (Table S9), which encodes for a protein that acts as an inhibitor of calcium/calmodulin-dependent protein kinase II (*camkII*). *camkII* has a vital role in long-term potentiation of synaptic strength (LTP) and learning, via regulation of glutamate receptors (AMPA)[59–63]. *camkII* is also one of the main calcium/calmodulin targets after the activation of NMDA (N-methyl-d-aspartate) glutamate receptors, which are involved in memory formation[64]. Moreover, a peptide derived from *camk2n2* (tatCN21) impairs fear memory formation by blocking *camk* activity[65], and overexpression of *camk2n2* in the hippocampus was found involved in memory formation[66]. In the Bengalese finch *Lonchura striata domestica*[67], one of the species included in the Cactus alignment, the glutamatergic system contributed to the attenuation of stress response and aggressive behaviour under domestication. Finally, in high stress lines of the domesticated Japanese quail *Coturnix japonica*, *camk2n2* and *camkII* have been detected as deleted, together with other genes in the same networks[68,69]. Loss of genes in this network may be responsible for the reduced growth rate and low basal weight of the high stress quails compared to low stress lines[69]. Since *camk2n2* is likely involved in behavioural and physiological changes under domestication in birds, we evaluated its conservation in relation to the onset of synanthropic habits in the barn swallow. We generated an alignment of transcripts from 38 species (17 domesticated or synanthropic, 21 wild; Table S31). However, we did not observe any pattern specific to domesticated or synanthropic species, and the single-gene phylogenetic tree substantially matched the known phylogeny. Thus, any role of *camk2n2* in synanthropic habits or domestication would have to be ascribed to non-coding regulatory elements. In vocal learning bird species, domestication was also found involved in the control of dopaminergic signalling in neural circuits that are crucial for vocal learning[67]. Among the top 20 genes with the most overlap between CDS and phyloP conserved bases (Table S9), *foxp2* has 74% of its CDS bases conserved. This gene received great attention for its role in language and speech, since mutations in its sequence cause, among others, speech impairments[70–75]. In the zebra finch, a vocal learner like the barn swallow, this gene has a marked expression in brain regions involved in song learning[76–79]. Another candidate gene detected

and previously associated with song learning is *ube2d3* (75% CDS conserved; Table S9), a gene located in a region of the human genome associated with musical abilities[80–82], which include recognizing, reproducing and memorising sounds. *camk2n2*, *foxp2* and *ube2d3* were also in the top 5% genes with the most overlaps between CDS and CEs bases detected with phastCons (Table S32).

The 5 top genes with the most overlaps between PhyloP conserved sites and CDS were stress-related genes (Table S9). In addition to *camk2n2* described above, *inhbb* is associated with pituitary hormones and its expression is affected by stress conditions (e.g.[83–85]). *sumo2* is involved in inflammatory and stress responses through its conjunction with *sumo3*. For instance, it promotes vascular oxidative stress in mice[86] and is responsible for protective stress responses to cope with stressful conditions in the brain[87]. *nfia* is another stress-related gene that was found involved in the onset of anxiety-like behaviours in adult mice after experiencing early-life stress[88]. Similarly to *nfia*, *sox2* deficiency in the murine suprachiasmatic nucleus (SCN), the central circadian clock, is involved in the perturbation of mood-associated phenotypes, such as anxiety- and depressive-like behaviours[89]. SCN neurons are GABAergic, and therefore involved in glutamatergic processes[90] such as *camk2n2*, the top candidate. Finally, *cnot2* was found involved in the maintenance for cell viability and its depletion induces a cascade that generates endoplasmic reticulum (ER) stress[91]. Among the cited genes, 3 of them are also involved in ER stress. Increased conjugation *sumo2/3* induces ER-stress-mediated cell death[92], *sfia* upregulation reduces the ER stress [93], *sox2* is downregulated when ER occurs[94]. They are therefore involved in cell-death pathways.

7 - SNP catalogue description

7 - 1 - SNP counts

A comprehensive SNP catalogue was generated using all publicly available genomic data for the barn swallow, including individuals from all extant subspecies as well as five high coverage samples sequenced with HiFi technology (Figure 3A). Genetic marker counts obtained after variant filtering (Table S21 and Methods) and average site depth and average individual depth for each dataset are reported in Table S33. Results relative to the HiFi dataset (for both full coverage and downsampled datasets) are discussed in more detail in Data S1 section 9-2. SNP counts and density for all chromosomes are summarised in Table S13. No SNP was detected on chromosome 39. For each sequencing technology, we also computed the number of SNPs by annotation category (genic, intergenic, exonic, intronic; Table S34).

7 - 2 - Correlation between SNP density and genomic features

For all datasets, SNP counts across all chromosomes were correlated with GC content by computing Spearman nonparametric rank test[95]. We found a positive correlation between chromosome GC content and SNP density (GC content was calculated as the fraction of GC bases every 10 kbp; SNP density was calculated over 10 kbp windows across all genome as described in Methods) in all datasets (HiFi WGS Spearman's $\rho = 0.25$, $S = 1.59 \times 10^{14}$, p-value $< 2.2 \times 10^{-16}$; downsampled 5x HiFi WGS Spearman's $\rho = 0.17$, $S = 1.75 \times$

10^{14}, p-value < 2.2 x 10^{-16}; Illumina WGS Spearman's ρ = 0.41, S = 1.25 x 10^{14}, p-value < 2.2 x 10^{-16}; Illumina ddRAD Spearman's ρ = 0.016, S = 2.08 x 10^{14}, p-value = 2.4 x 10^{-7}).

8 - Marker catalogue applications

8 - 1 - Relationship between LD and distance from chromosome ends

Different studies have investigated the relationship between LD values and distance from chromosome ends, also evaluating the presence of stable recombination hotspots across chromosomes[96–98]. A significant increase in recombination rate was detected towards chromosome ends in the collared flycatcher genome[96]. Thus, after finding that microchromosomes exhibit lower LD than macrochromosomes (Figure 4B), we used the SNPs in our catalogue from Illumina WGS ds2.1, 2.2, 3.1.1, 3.1.2 to evaluate a potential correlation between LD and distance from chromosome ends, according to the different chromosome types (Figure S11). We detected a positive correlation between distance from chromosome ends and LD values for macrochromosomes and intermediate chromosomes in all datasets, except for *H. r. rustica* (ds2.2) macrochromosomes, where the correlation is negative (Table S35). Concerning microchromosomes, we only detected a significant positive correlation in ds2.1 and a significant negative correlation in ds2.2. A positive correlation between distance from chromosome ends and LD would be consistent with the increase in recombination rate observed towards chromosome ends in the collared flycatcher[96]. A detailed analysis regarding the heterogeneity of the recombination landscape along the barn swallow genome is warranted to gain better insights about the different result obtained in ds2.2.

8 - 2 - Genome-wide scans

8 - 2 - 1 - Genes in high LD blocks

LD reflects the evolutionary history of populations as it can be influenced by selective pressures[99–101], recombination rate[98,102], migration[103], genetic drift[104] and population admixture[105,106]. Thus, to further exploit our genetic markers catalogue and generate a list of potential candidate genes under selection to confirm with future studies in this species, we performed an initial chromosome scan using Illumina WGS data from the *H. r. erythrogaster* and *H. r. savignii* subspecies[107] (ds3.1) to identify potential regions of interest (ROIs) exhibiting high LD values (average $r^2 > 0.3$). Despite the small sample size and the rapid genome-wide LD decay, our analyses revealed the presence of 78 ROIs, many of which (n = 57/78) span at least one annotated protein coding gene (N = 83; Table S22). Excluding ROIs containing sequences potentially collapsed in the reference or not overlapping with annotated genes, the locus showing the highest r^2 values is on chr6 (ROI 45) and harbours four genes (*ccdc34*, *lgr4*, *lin7c* and *bdnf*; Figure S9a, Table S22). Among these, *lgr4* and *bdnf*, due to their well-documented role also in birds, can be considered particularly interesting for future studies on the species. *bdnf* encodes a major neurotrophin involved in neuronal plasticity and differentiation[108,109]. In zebra finch males, its transcript is upregulated to high levels in the high vocal centre (HVC) by singing activity[110], particularly when juveniles start to emit vocalisations, and its tissue-specific overexpression significantly increases during sensorimotor song learning[111–113]. *Bdnf* is also implicated in neural

crest cells development[114], and studies in multiple domesticated mammalian species suggest a role for the modification of neural crest development in driving the concerted evolution of tame phenotypes during domestication (i.e., 'domestication syndrome')[115,116]. Moreover, it is extensively implicated in the response to stress, fear, and fear memory consolidation[117]. Similarly to other species[118], barn swallow *bdnf* presents alternative transcripts (Figure S9B), three of which (transcript variants X2, X3, X4) lead to the same amino acid sequence, suggesting the presence of important regulatory elements. In other bird species, temperature (chicken[119]) and prolonged social isolation (zebra finch[120]) affect the expression of *bdnf* through a methylation-mediated mechanism associated with CpG sites located within CpG islands upstream of the translation start site, as well as in the coding region. Initially, using WGS data from American and Egyptian samples[107] (ds3.1), we detected 6 LD blocks comprising 104 SNPs within the *bdnf* gene region (Figure S9B). Of these SNPs, 30 directly alter CpG sites, either in the reference or in the alternate allele sequence (Table S36). The highest LD values were identified within *H. r. savignii* population (Figure S9D), where we also detected an average homozygosity (i.e. the average proportion of homozygous genotypes) of ~88.8% across all samples for the genotyped SNPs within the gene (Table S36). The same genomic region in all other available WGS populations (ds2) has similar LD patterns (Figure S9D). For instance, *H. r. transitiva* shows very high pairwise LD values within *bdnf* gene coordinates (Figure S9D). Four CpG islands are present within the sequence of *bdnf* in the barn swallow (Figure S9B, blue blocks). The first CpG island corresponds to one of the two genomic regions containing methylated sites previously described in zebra finch[120]. We found that four of the seven CpG sites reported in zebra finch are conserved in the barn swallow (Figure S9C, highlighted in yellow). One SNP present in our barn swallow markers catalogue (chr6:53,908,036) directly affects a CpG site adjacent to a zebra finch methylation site[120] (Figure S9C, SNP adjacent to the first highlighted CpG site). We also analysed this region in the Cactus multialignment and found that all the zebra finch CpG sites are conserved in all other bird species, except for the chicken, where only two sites are conserved as CpG (Figure S9C). The presence and conservation of CpG sites in the barn swallow reinforce the importance of these sites. CpG islands are known to directly affect the transcription of genes by altering local chromatin structure, mostly through methylation of CpG dinucleotides[119]. For *bdnf*, methylation-dependent transcriptional regulation involving CpG islands has been shown to affect fear memory consolidation[121], a process strictly involved in domestication. After confirming a selective pressure acting upon this gene, methylation state assays could help to further investigate the role played by epigenetic modifications of *bdnf* in the barn swallow. Due to *bdnf* well-documented role in stress response and fear memory consolidation, an intriguing possibility could be that the strict association with humans in this species is linked with the evolution of pathways suppressing fear response and promoting tameness, that are typically under selection in domesticated taxa.

lgr4 (Leucine-rich repeat-containing G protein-coupled receptor 4) encodes a G protein coupled receptor well studied for its role in modulating cells responsiveness to Wnt ligands[122] and in regulating energy metabolism, including food intake, energy expenditure and lipid metabolism[123]. In yellow-feathered chicken breeds, through genome-wide scans for selective sweeps and runs of homozygosity analysis, *lgr4* was identified as one of the major candidate pigments determining genes[124,125]. Previous studies correlated the function of this gene to pigment deposition[126] and hair follicle development[127]. Moreover, in a recent genome-wide run of homozygosity analysis conducted on Tibetan native chickens a genomic interval harbouring *bdnf*, *ccdc34*, *lgr4*, *lin7c*, *gls*, *loc101747789*, *myo1b*, *stat1* and

stat4 stood out as candidate region for essential roles in adaptation of this species[128]. Notably, four of the genes present in this region (*bdnf*, *ccdc34*, *lgr4*, *lin7c*) are the genes constituting the high-LD ROI 45 in our barn swallow marker catalogue.

8 - 2 - 2 - Population haplotype homozygosity statistics scans

Since a genome-wide LD scan alone is not sufficient to accurately identify candidate genes, to try to confirm the presence of a potential selective signature within this genomic region (ROI 45), we computed population haplotype homozygosity statistics (iHS, the integrated haplotype homozygosity score) on chr6 in WGS ds3.1, ds2.1 and ds2.2. The ROI harbouring *bdnf* and *lgr4* identified with genome-wide LD scans is associated with significant outlier peaks also after this analysis (Figure S12). Yet, we are aware that these results might not be considered as definitive, due to the limited sample size of the available populations and the only partial phasing achievable with these short-read genomic data. Even if these two analyses (LD and iHS scans) at the present time are not sufficient to conclusively state these genomic regions are under selection in the barn swallow, for the limitations underlined above, the described genes very likely represent interesting candidates to focus for future research. One potential confounding factor to exclude is that the low diversity detected in these genomics regions might be partially due to low recombination rate patterns rather than selective pressure only[129]. Hence, complementing these results with the reconstruction of the recombination landscape along the barn swallow genome will provide useful insights for better evaluating the unusual selective pressure in this genomic region suggested by our analyses.

9 - HiFi read mapping, variant calling, titration and phasing experiments

9 - 1 Sequencing and read mapping

HiFi WGS for the five samples from ds1 generated a mean read length across samples of 15609.7 bp (≥ Q20). About 25.2 Gbp of ≥ Q20 data were generated on average for each sample, and median read ≥ Q20 quality averaged across samples is 31 (Table S11b). Read alignment has a mean mapped concordance of ~97% across all samples, with an average of 1.64 million reads mapped to bHirRus1 (Table S11g). We verified read alignment coverage for each sample using bedtools genomecov and verified that the distribution of bases at a certain coverage value is consistent with the expectation based on sequencing coverage (Figure S13A-S13E). The secondary peak observed at half average coverage, particularly in samples sequenced at higher coverage (A1, 2, 4), is presumably due to the fact that all five samples belong to the heterogametic sex. We also compared the proportion of the genome covered by uniquely mapped reads between Illumina WGS reads (96.35%) and HiFi reads (97.25%).

9 - 2 Variant calling and titration experiment

Given the high accuracy of the HiFi reads, we performed a titration experiment to ascertain how sequencing depth affects variant calling. This aimed at identifying the minimum coverage required to detect reliable variants, with the goal of a recall rate > 60%. Recall rates from individual variant calling are reported in Table S37. We found that the steepest decrease in the number of variants detected occurs

between 7.5x and 5x (Figure S13F). At 5x, for all truth sets considered, ~50% of the variants called at 20x were not recalled (Table S37a). These preliminary results suggest sequencing coverage of about 7.5x to recover an adequate number of SNPs for subsequent analyses. Precision remains high even at 5x (average precision rate across all samples of 0.93 for truth set 1). Truth sets 2 and 3 were excluded from further evaluation because they resulted in very low values of precision, even at high coverage (Table S37a; this is presumably due to the fact that both these truth sets include only a limited subset of the variants present in our datasets). In addition, we performed joint variant calling to assess whether an increased number of variants can be recovered from the five combined HiFi read sets, by combining reads for variants shared between samples for a joint call, with subsequent genotyping for each sample for the variant. While joint calling for SNVs and indels was comparable to performance with individually called variants for full and downsampled coverage, the recovery of structural variants (including deletions, insertions, inversions, duplications, translocations, copy number variations) is significantly improved by joint calling compared to per-sample variant calling (72,691 structural variants on average per sample with joint calling, vs. 50,502 with per-sample calling, compared to 109,248 for full HiFi read set and joint calling; recall rate 0.665 vs. 0.462; Figure S13G-S13I; Table S37b). These data suggest that, thanks to the significantly higher variants recall achieved with a joint call approach, lowering the sequencing coverage to 5x can be considered as an alternative approach when working with HiFi data, particularly when interested in recovering structural variants.

9 - 3 Haplotype phasing with HiFi reads

We took advantage of the high accuracy and length of HiFi reads to reconstruct phased haplotype blocks across chromosomes for all 5 samples (Table S38). The percentage of phased blocks across chromosomes for sample A1 is shown in Figure S14A. We obtained 73% of average phased sequence for macrochromosomes, 77% for intermediate and 64% for microchromosomes (Figure S14B). Phasing in sex chromosomes is lower than 6%, due to all samples being females, which can be regarded the average error of the method. This can be caused by regions collapsed in the assembly or by regions with high repeat content.

Supplementary Figures

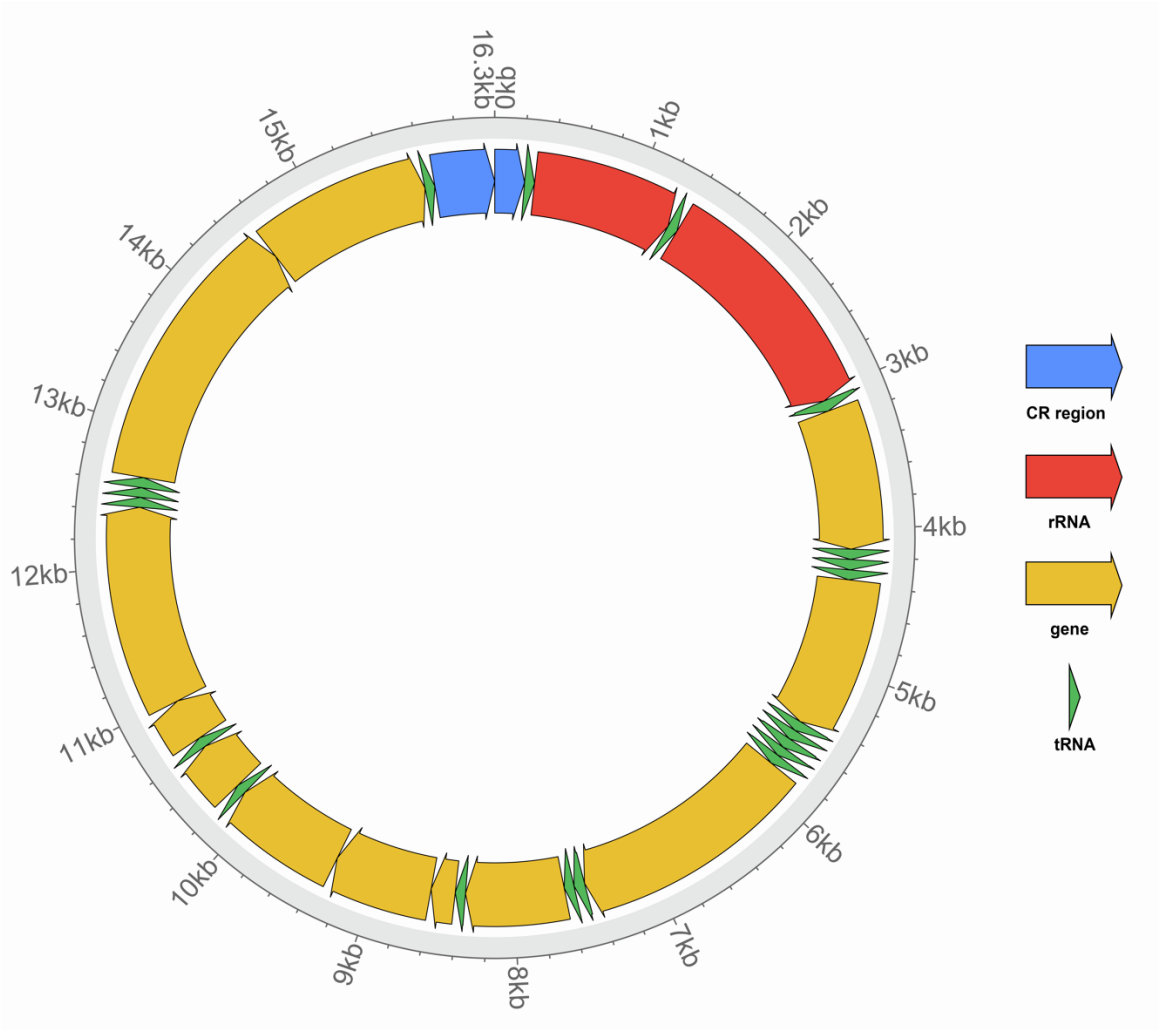


Figure S1. Circular representation of the barn swallow mitogenome assembled with NOVOplasty[130]. Related to Data S1.
The control regions (CR) are coloured in blue, rRNA genes in red, tRNA genes in green, other genes in yellow.

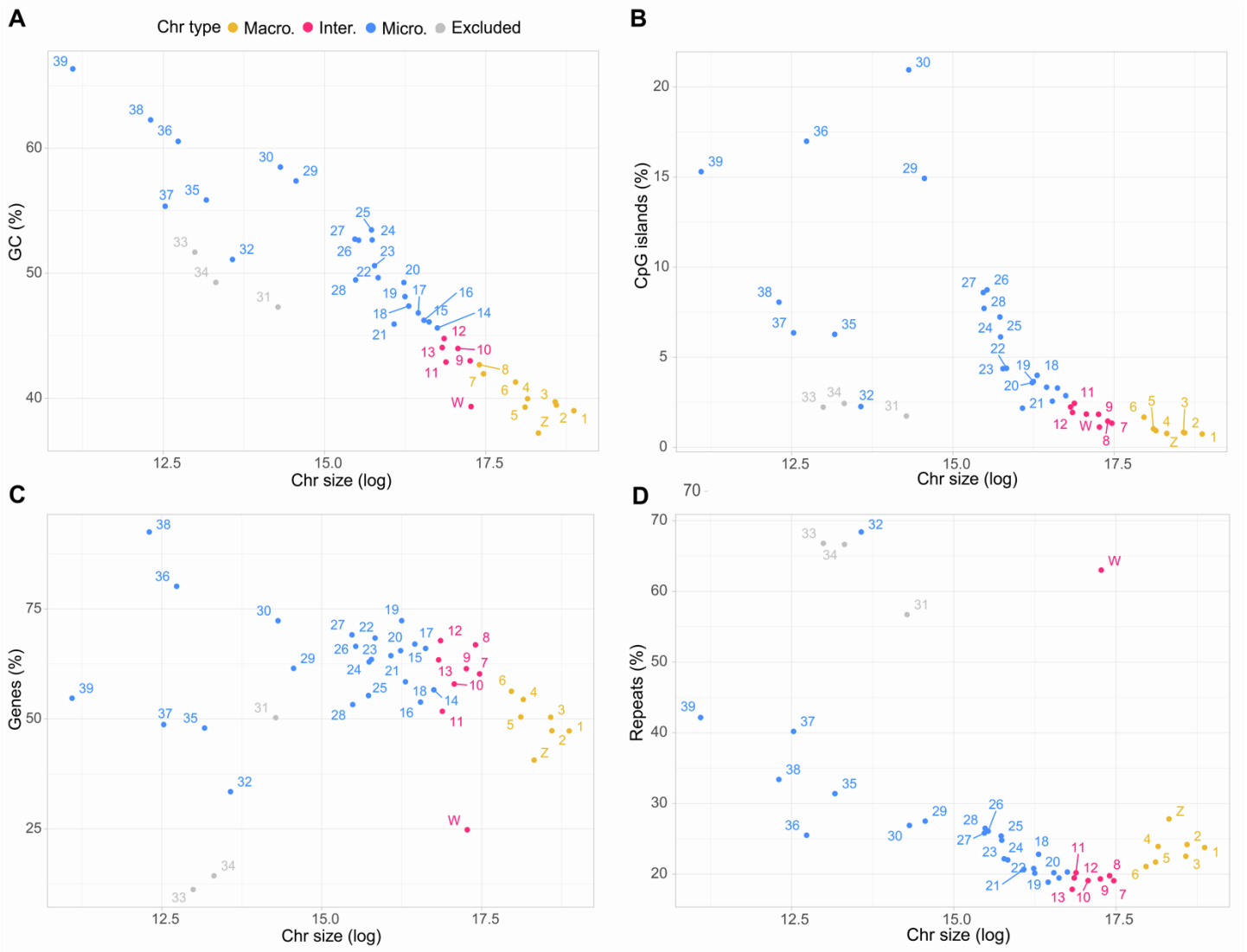


Figure S2. Correlation between chromosome length (log) and genomic features. Related to Figure 2C.
Chr 31, 33, 34 were excluded based on their higher PacBio raw-reads coverage with respect to the other chromosomes. **(A)** GC content **(B)** CpG islands. **(C)** Gene content. **(D)** Repeat content.

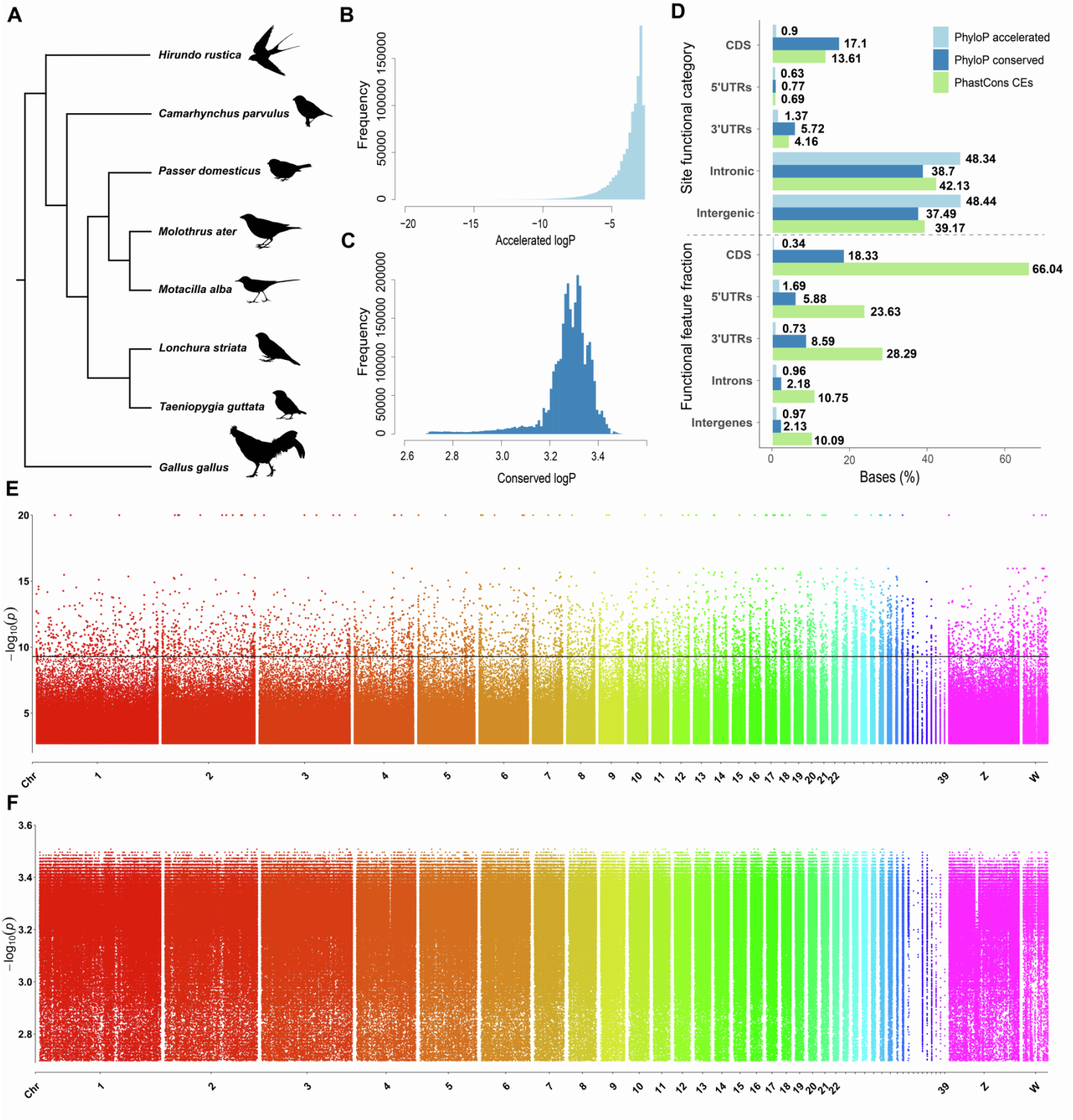


Figure S3. Cactus alignment and selection analysis. Related to Figure 2C.
(A) Cactus aligned species tree topology. Cactus guide tree: (Gallus_gallus:98.04286929, ((Hirundo_rustica:43.70000000, (Ficedula_albicollis:42.77829604,Parus_major:42.77829604):0.92170396):0.00000000, (Camarhynchus_parvulus:38.00000000, ((Passer_domesticus:34.80000000, (Molothrus_ater:24.06782500,Motacilla_alba:24.06782500):10.73217500):0.00000000, (Lonchura_striata:10.12359500,Taeniopygia_guttata:10.12359500):24.67640500):3.20000000):5.70000000):54.34286929);. **(B)** Frequency distribution of accelerated sites logP values. **(C)** Frequency distribution of conserved sites logP values. **(D)** Intersection between PhyloP conserved and accelerated, and PhastCons CEs bases with bHirRus1 genomic features. Top panel represents the percentage of PhyloP/PhastCons bases that overlap with genomic features, while the bottom panel represents the fraction of genomic features covered by accelerated or conserved bases. The CDS were the most conserved according to both phyloP and phastCons. **(E)** Manhattan plot of the FDR-corrected accelerated sites identified with PhyloP. Each point represents a 10 bp conserved or accelerated site. LogP values are on the x axis. The horizontal line defines the Bonferroni threshold (9.3). **(F)** Manhattan plot of the FDR-corrected conserved sites found with PhyloP. No significant bases were found after Bonferroni correction.

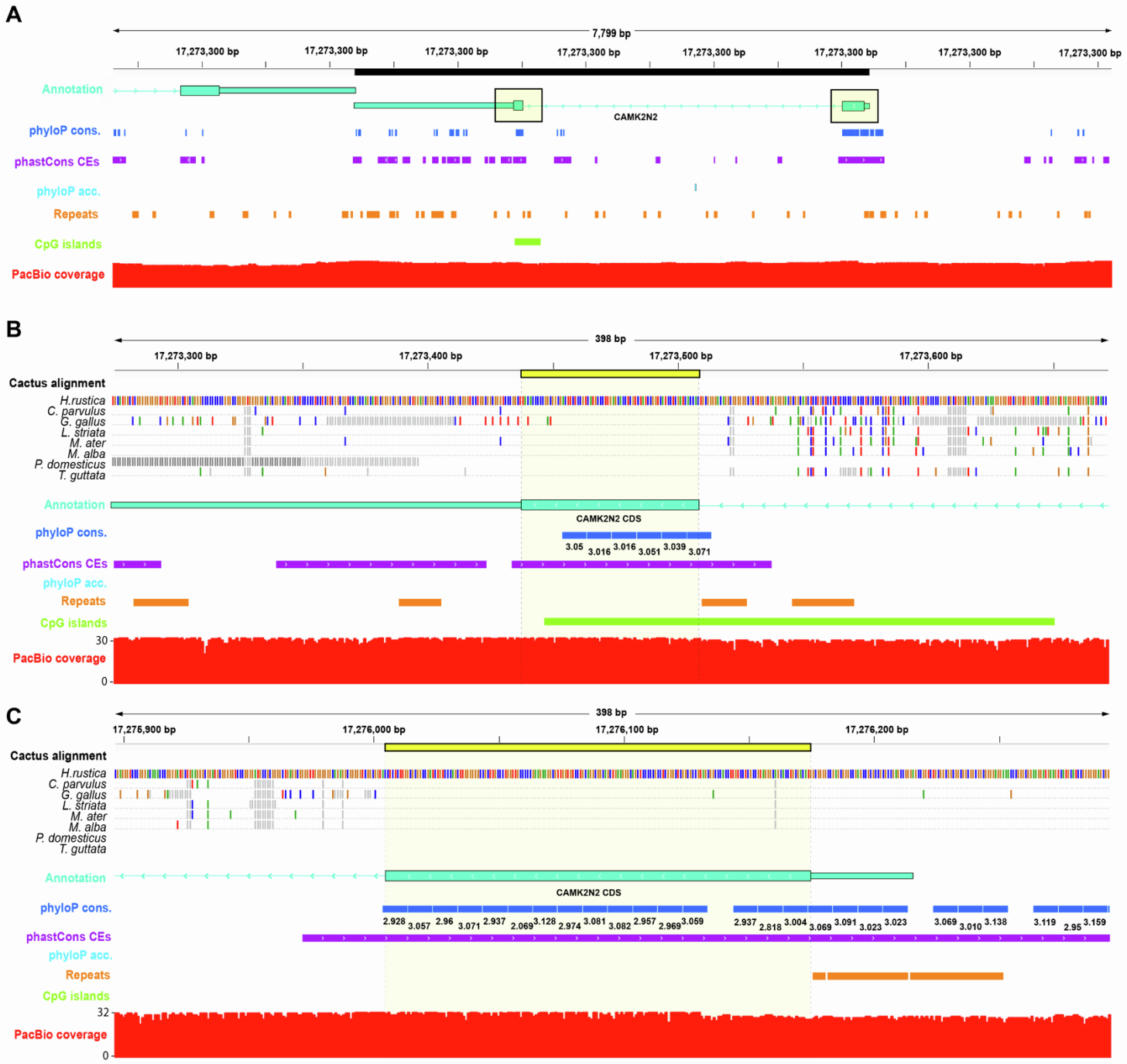


Figure S4. CDS of *camk2n2* on chr10 visualised with IGV[131]. Related to Figure 5F.
(A) Whole gene. The black bar represents the gene coordinates. The yellow rectangles highlight the CDS regions. **(B)** First CDS (17,273,437-17,273,508). The yellow horizontal bar represents the CDS coordinates. Tracks represent Cactus MAF alignment referenced to the barn swallow, gff3 bHirRus1 gene annotation, phyloP conserved 10 bp windows with their logPvalue, phastCons CEs, PhyloP accelerated 10 bp windows, repeats, CpG islands and PacBio coverage. In the Cactus alignment track, all the barn swallow bases and divergent sites in the alignment are shown (green A, red T, blue C, orange G). Dots represent identical sites. SNPs in the other species are represented with bases with different colours than the barn swallow, while the points represent sites with the same base calling. It can be noticed from the picture how phastCons CEs are more inclusive than phyloP sites, which instead are more precise. **(C)** Second CDS (17,276,005-17,276,174).

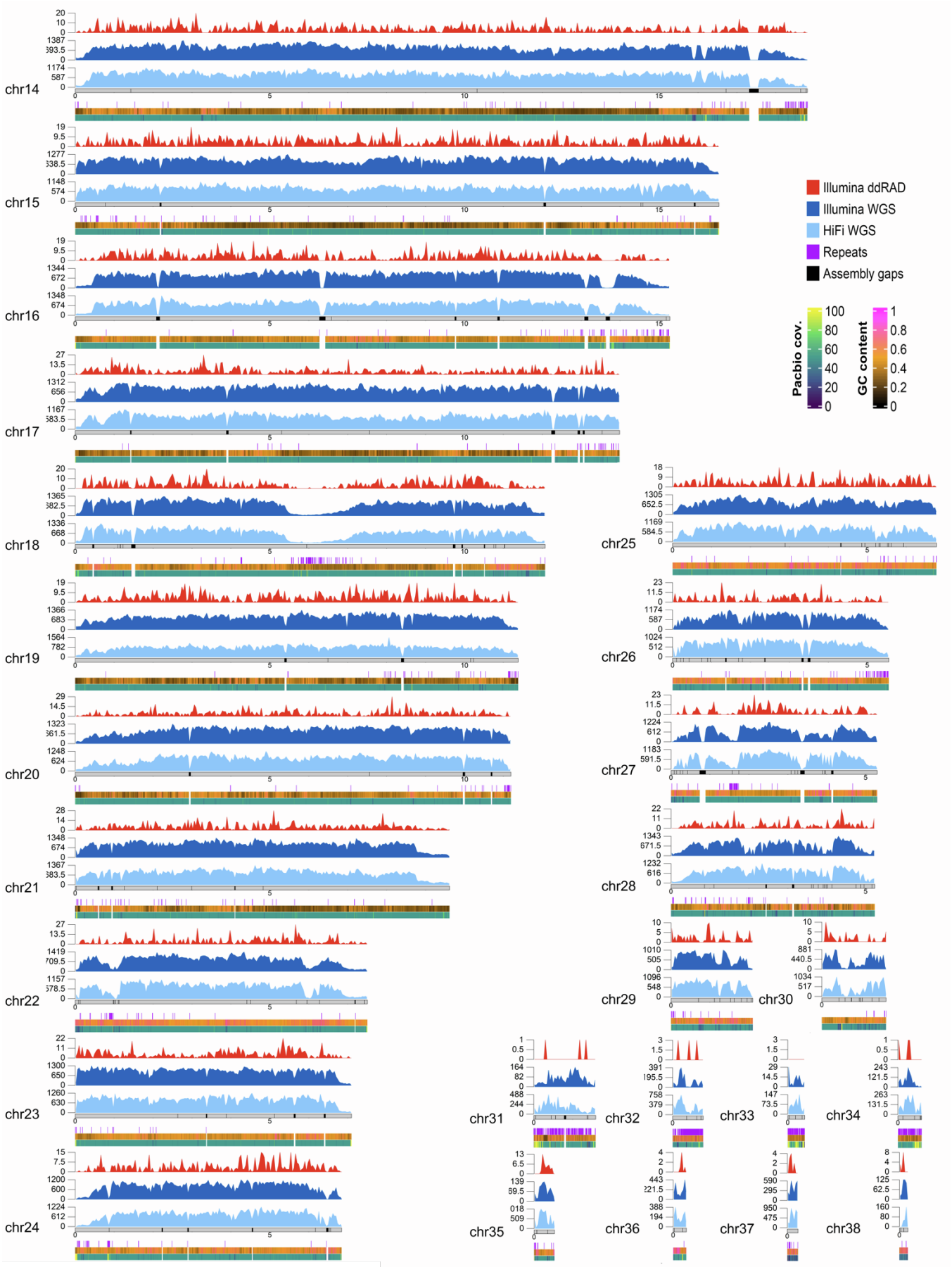


Figure S5. SNPs density across microchromosomes. Related to Figure 3B.
SNP density, coloured according to the different types of genomic data used, was computed over 40 kbp windows. The numbers on the y axis of each density track indicate the maximum and average values of SNP density for each track. Light blue: HiFi WGS data (ds1). Dark blue: Illumina WGS data from ds2 and ds3.1. Red: Illumina ddRAD data from ds3.2 through ds6.8. All available samples from the same sequencing technology were considered together. Additional tracks in the lower panel show repetitive regions of the genome (violet bars; only regions larger than 1 kbp are plotted), GC content and PacBio reads coverage. Grey ideograms represent chromosomes in scale, with assembly gaps highlighted as black bars.

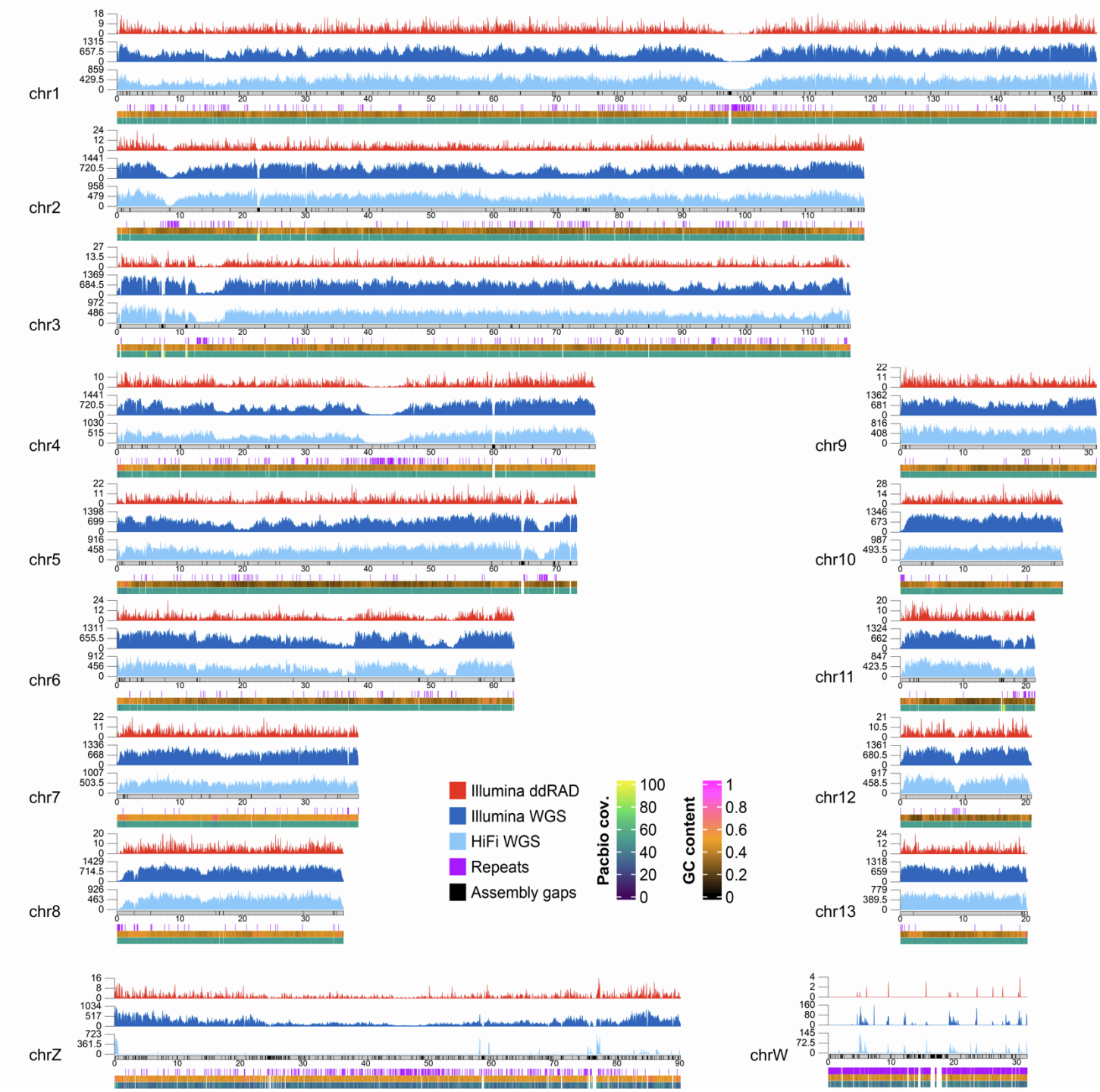


Figure S6. SNPs density per chromosome using a downsampled (5x) set of HiFi reads. Related to Figure 3B and Data S1.
Only macrochromosomes and intermediate chromosomes are shown. Sex chromosomes are shown at the bottom. SNP density, coloured according to the different types of genomic data used, was computed over 40 kbp windows. Maximum and average SNP density are indicated on the y axis. Light blue: HiFi WGS data (ds1). Dark blue: Illumina WGS data from ds2 and ds3.1. Red: Illumina ddRAD data from ds3.2 through ds6.8. All available samples from the same sequencing technology were considered together. Additional tracks in the lower panel show repetitive regions of the genome (violet bars; only regions larger than 3 kbp are plotted), GC content and PacBio reads coverage. Grey ideograms represent chromosomes in scale, with assembly gaps highlighted as black bars. The distribution of SNPs in the HiFi dataset after titration appears comparable to the full coverage distribution, as expected from a random subsampling of genomic data. The only chromosomes where downsampling led to a visible reduction in SNP density are sex chromosomes, consistent with their haploid coverage.

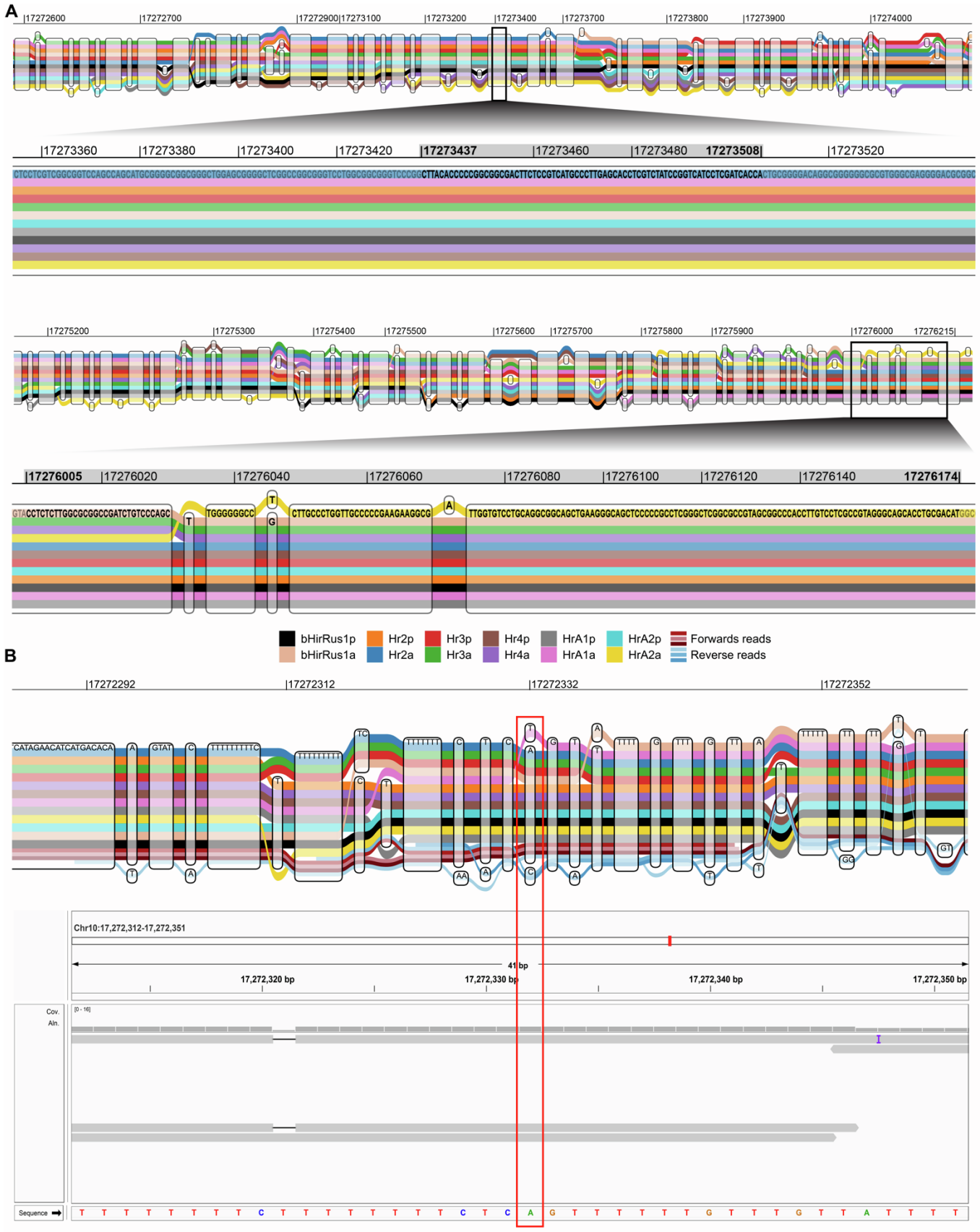


Figure S7. Graphical representation of the barn swallow's pangenome graph and variant calling comparison. Related to Figure 5F. **(A)** The top panel represents an example of visualisation of the initial region of *camk2n2*, the candidate gene with the most conserved CDS according to the comparative genomics analyses. The zoomed part shows the first CDS (grey rectangle, 17,273,437-17,273,508), which is highly conserved also between individuals of the same species. The bottom panel represents *camk2n2* terminal region. The zoomed part shows the details of the second CDS (grey rectangle, 17,276,005-17,276,174), which is also conserved. **(B)** The top panel is a graphical representation of the pangenome graph with the 16 Illumina WGS barn swallow individuals mapped on it. Aligned reads are represented by thinner lines (red: forward reads, blue: reverse reads). The bottom panel is a graphical representation of the alignment between the same 16 Illumina WGS reads and the linear reference genome (bHirRus1). The red square points at the SNP at position 17,272,332 within *camk2n2* region, which was called from raw reads aligned to the pangenome (top panel) and was not supported by reads aligned to the linear reference genome (bottom panel).

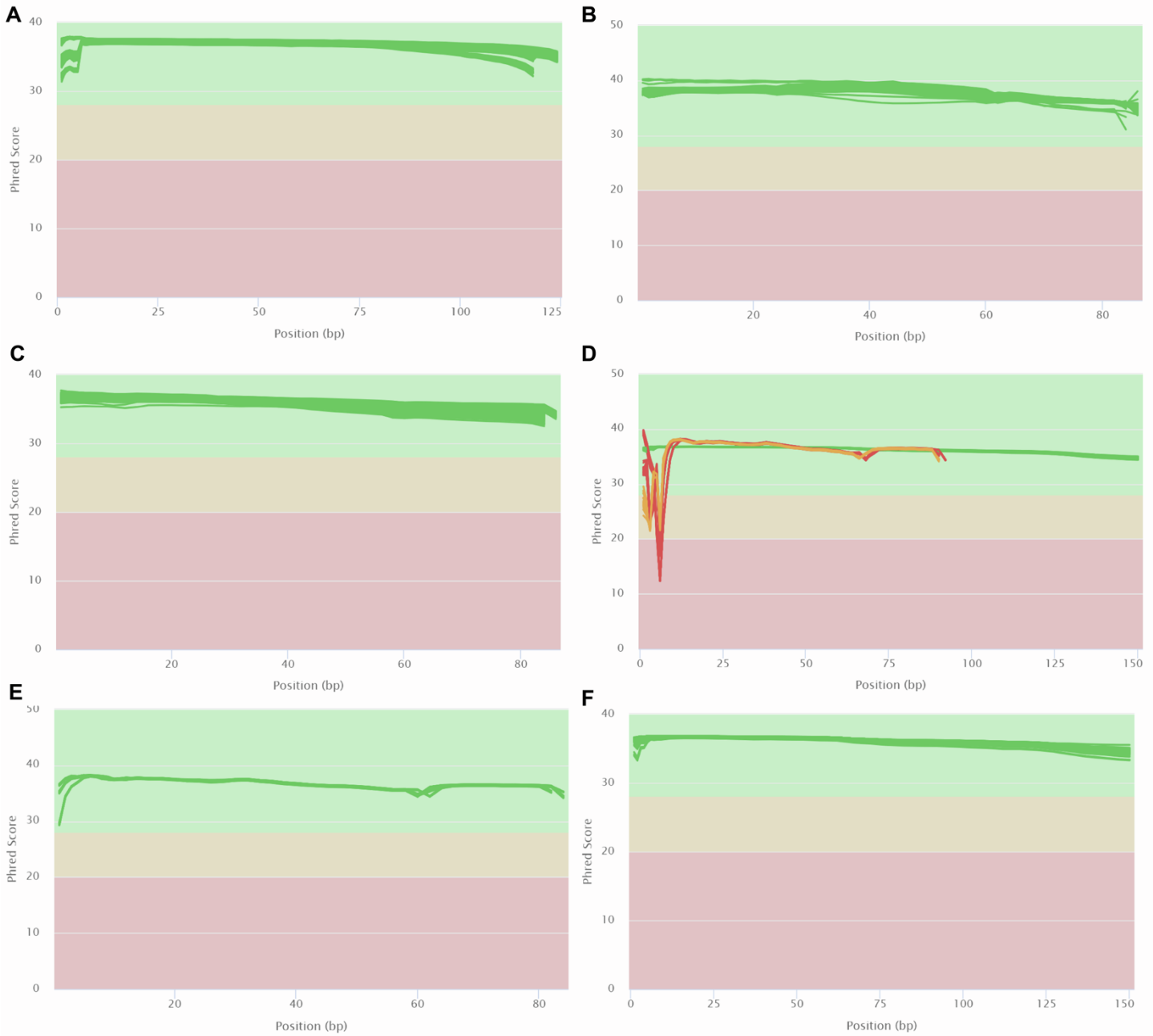


Figure S8. Per base quality scores (y-axis) from fastqc output performed on genomic data of the different datasets after adapter sequences trimming. Related to STAR Methods.

(A) ds5. **(B)** ds6. **(C)** ds4. **(D)** ds3. American and Egyptian samples (ds3.2.1 and ds3.2.2) showed low quality scores at the beginning of the reads and were further processed to trim low-quality bases. **(E)** ds3.2.1 and ds3.2.2 after removal of low-quality bases. **(F)** ds2.

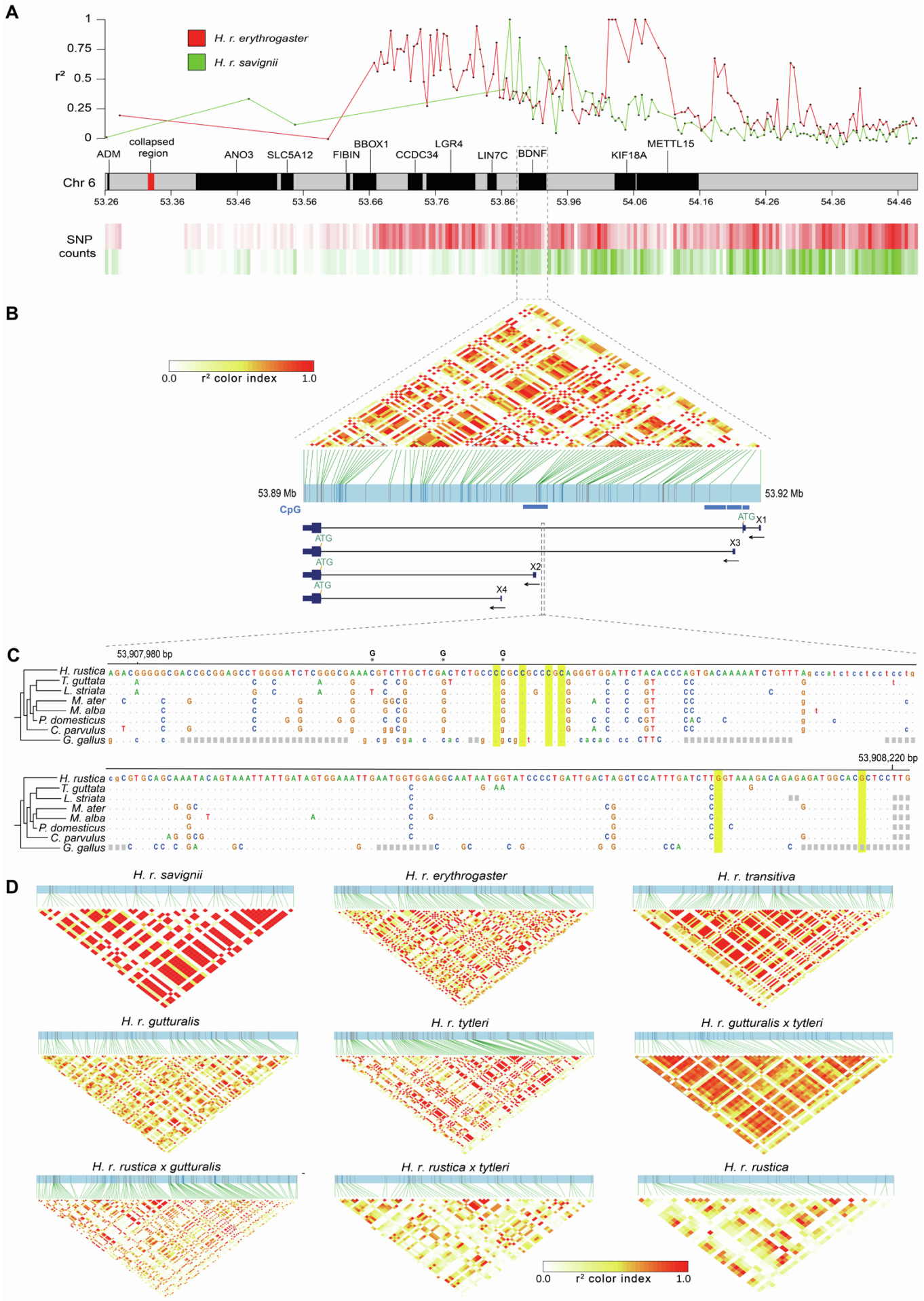


Figure S9. Patterns of LD blocks in genomic regions on chr6. Related to Figure 4 and Data S1.

(A) Average r^2 values computed over 5 kbp windows on chr. 6 (upper panel; from 53.26 Mb to 54.49 Mb) for the *H. r. savignii* (green) and *H. r. erythrogaster* (red) populations (ds3.1). The region shown in the plot extends beyond ROI 45. Each point represents the average r^2 value per window and was placed at the midpoint of the genomic region. The heatmap in the lower panel represents SNP counts for the two populations analysed. **(B)** Upper panel: LD heatmap within *bdnf* gene coordinates considering the two populations combined. Black triangles indicate LD blocks. Blue horizontal blocks mark the presence of CpG islands. Lower panel: barn swallow *bdnf* four transcript isoforms X1, X2, X3 and X4 (big rectangles: coding exons; small rectangles: noncoding exons; horizontal line: introns; arrows indicate the direction of transcription). **(C)** Cactus multiple alignment of the zebra finch (second line) region containing CpG sites important for methylation-dependent regulation[120]. Asterisks: SNPs present in barn swallow marker catalogue. Alternate base is shown on top of the barn swallow reference sequence. Yellow: zebra finch methylated sites[120]. The second, third and sixth CpG sites are conserved in the barn swallow. The first one (at position 53,908,035) is not fixed in the barn swallow but the transition of the adjacent polymorphic site from reference (C) to alternate (G) allele leads to the formation of a CpG site. **(D)** LD heatmap within *bdnf* gene coordinates (chr6: 53,886,627-53,927,580) in Illumina WGS from dataset2 and dataset3.1.1. *H. r. savignii* population (ds3.1.1). *H. r. erythrogaster* population (ds3.1.2). *H. r. transitiva* population (ds2.3). *H. r. gutturalis* population (ds2.1). *H. r. tytleri* population (ds2.4). *H. r. gutturalis x tytleri* population (ds2.5). *H. r. rustica x gutturalis* population (ds2.6). *H. r. rustica x tytleri* population (ds2.7). *H. r. rustica* population (ds2.2).

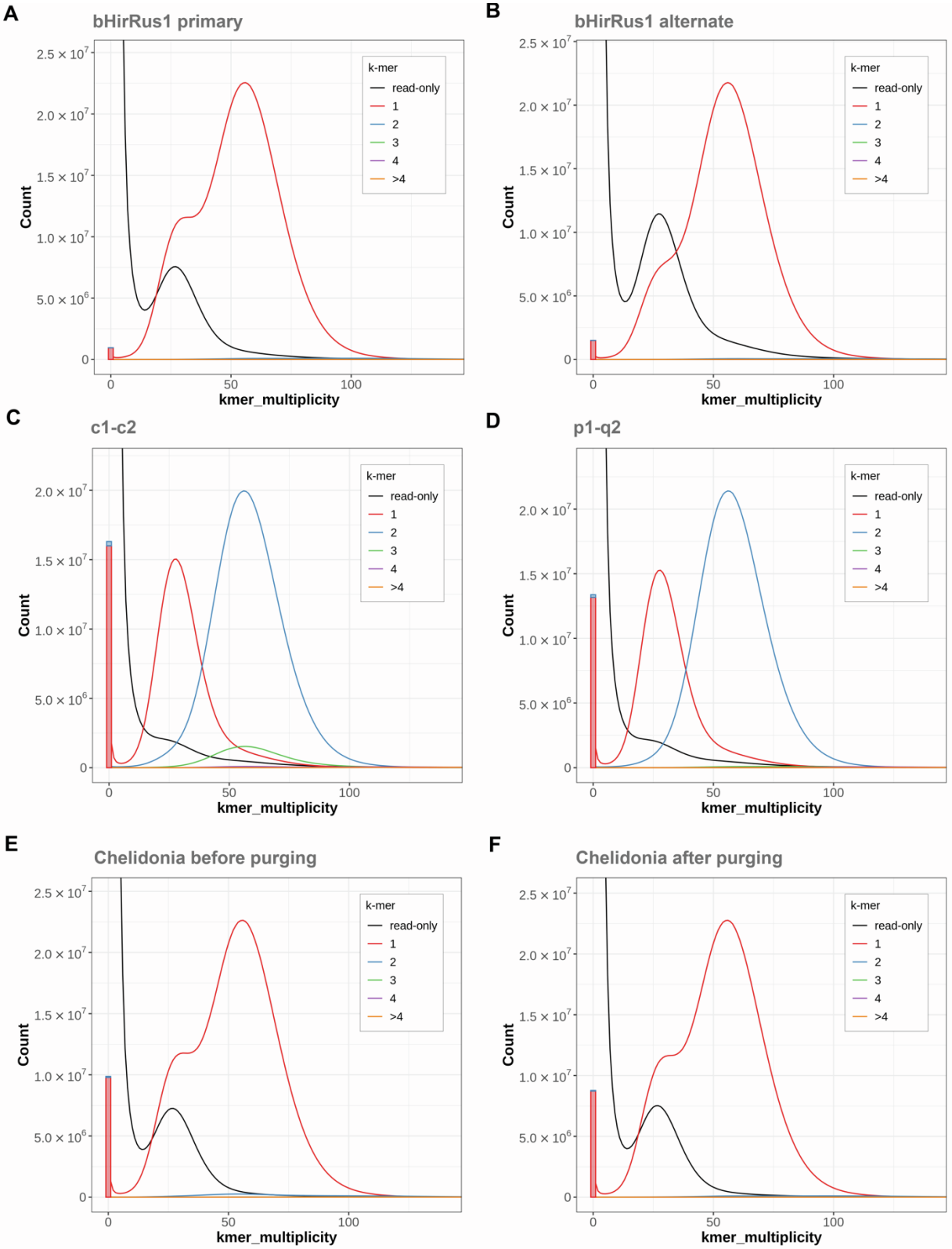


Figure S10. Merqury unstacked spectra-cn plots. Related to Figure 1 and Data S1.
(A) bHirRus1 primary assembly only. **(B)** bHirRus1 alternate assembly only. **(C)** Initial pseudo-haplotype assembly including primary contigs (c1) and alternate haplotigs (c2) generated by FALCON-unzip before purging. Retained haplotigs in the primary assembly show up as 3-copy *k*-mers (green curve) at about 50x (diploid coverage). **(D)** Pseudo-haplotype assembly after purge_dups[6] (p1 and q2 intermediates). Purging effectively removed 3-copy *k*-mers. **(E)** Chelidonia haploid assembly shows a higher number of false *k*-mers (bars at the origin) compared to bHirRus1 (panel A), as well as the presence of *k*-mers with diploid coverage in the read set (~50×) found twice in the assembly (blue curve). **(F)** Chelidonia after purge_dups[6]. Duplicated *k*-mers were removed (blue curve reduced).

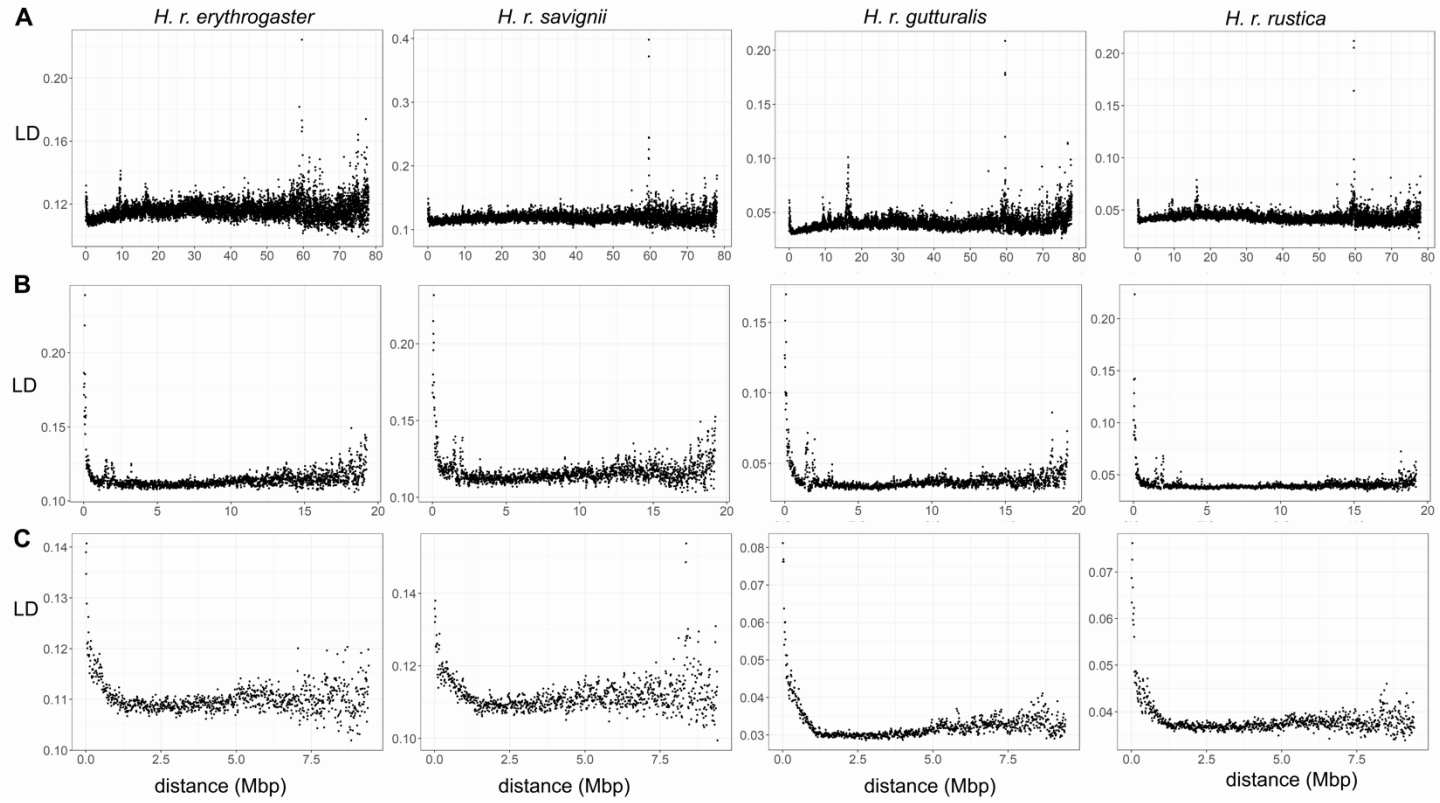


Figure S11 - Relationship between LD and distance from chromosome ends. Related to Data S1.
The analysis was run for different subspecies: *H.r.erythrogaster* (ds3.1.2). *H.r.savignii* **(**ds3.1.1). *H.r.gutturalis* (ds2.1). *H.r.rustica* (ds2.2). Average LD (r^2) values were computed by grouping marker pairs in 10kb distance (from both chromosomes ends) bins, divided according to chromosome type. **(A)** Macrochromosomes. **(B)** Intermediate chromosomes. **(C)** Microchromosomes.

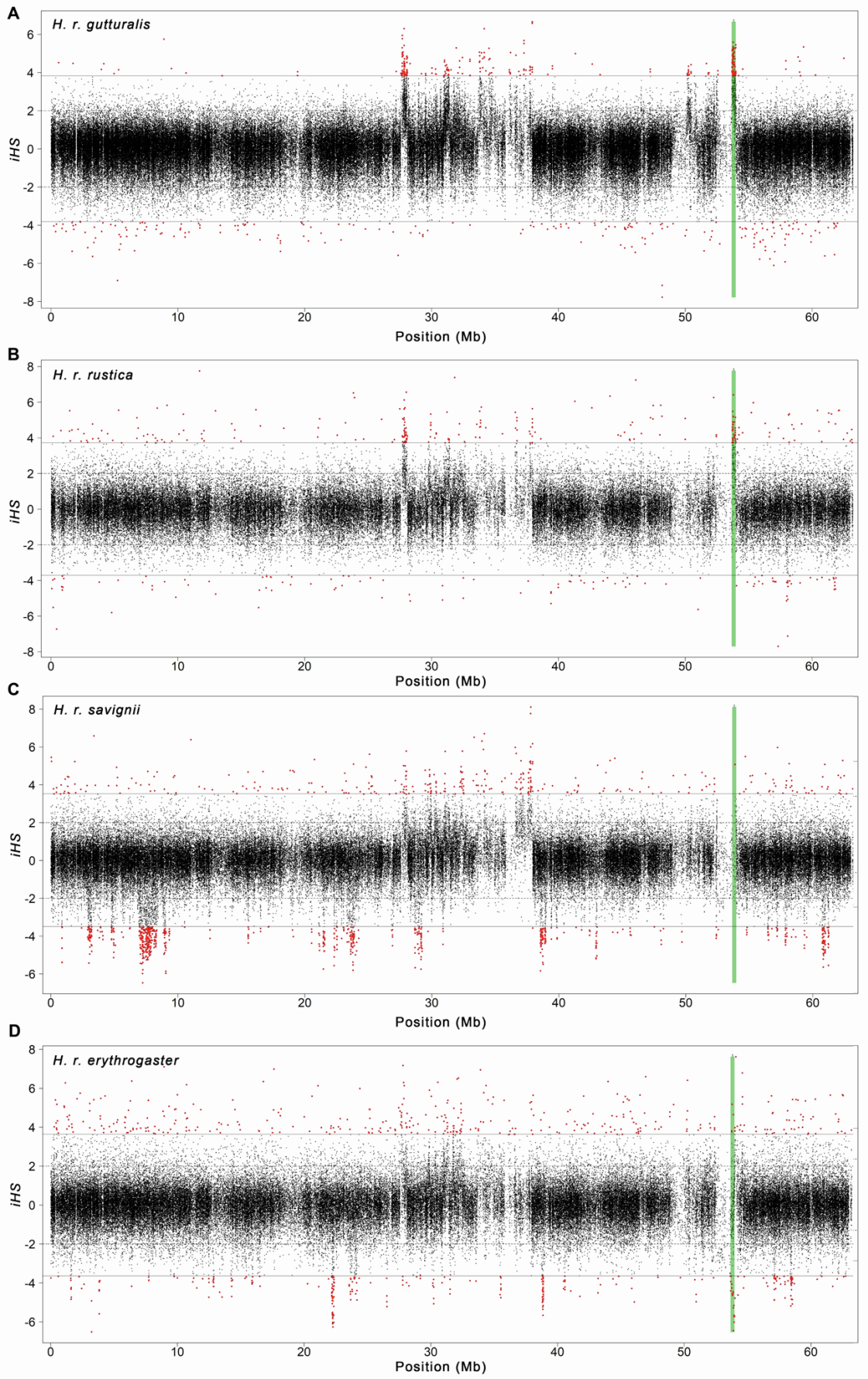


Figure S12. Integrated haplotype homozygosity score (iHS) computed on chr6. Related to Data S1.
(A) *H. r. gutturalis* population (ds2.1). Horizontal dashed lines (at +2, -2 values) represent the threshold identifying statistically significant iHS scores. Horizontal solid lines represent the threshold calculated after FDR correction. Statistically significant values after FDR correction are highlighted in red. The green vertical bars indicate the ROI with high LD values harbouring the *bdnf* gene (chr6: 53,680,000-53,954,999). **(B)** *H. r. rustica* population (ds2.2). **(C)** *H. r. savignii* population (ds3.1.1). **(D)** *H. r. erythrogaster* population (ds3.1.2).

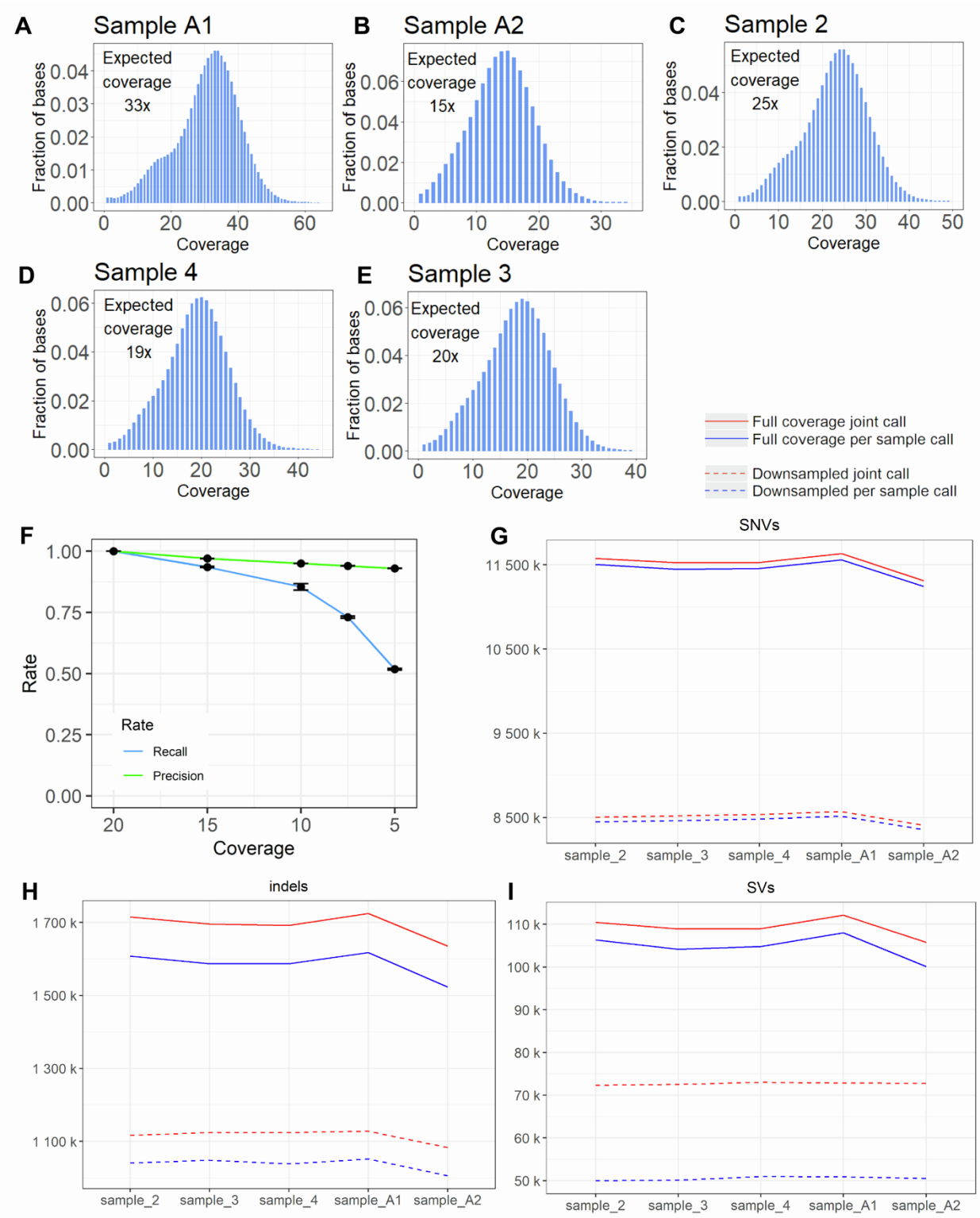


Figure S13. Histogram of alignment coverage for ds1 samples, recall and precision rates in individual titration experiments at varying HiFi coverage (x-axis) and absolute counts of joint and solo variant calls for the five HiFi read sets using the full or downsampled (5x) data sets. Related to Data S1. (A) Histogram of alignment coverage for samples A1 (ds1). Per-base coverage was computed with bedtools genomecov. **(B)** Histogram of alignment coverage for samples A2 (ds1). Per-base coverage was computed with bedtools genomecov. **(C)** Histogram of alignment coverage for samples 2 (ds1). Per-base coverage was computed with bedtools genomecov. **(D)** Histogram of alignment coverage for samples 4 (ds1). Per-base coverage was computed with bedtools genomecov. **(E)** Histogram of alignment coverage for samples 3 (ds1). Per-base coverage was computed with bedtools genomecov. (F) Recall and precision rates in individual titration experiments at varying HiFi coverage (x-axis). Each observation represents the average rate with standard deviation computed across all five samples (ds1) considering the full read set as truth. Recall rate appears to sharply decrease between 10x and 5x, while precision rate does not decrease significantly at low coverage values. **(G)** Absolute counts of joint and solo variant calls for SNVs for the five HiFi read sets using the full or downsampled (5x) data sets. **(H)** Absolute counts of joint and solo variant calls for indels (per-sample calls from DeepVariant, joint calls from DeepVariant and GLNexus) for the five HiFi read sets using the full or downsampled (5x) data sets. **(I)** Absolute counts of joint and solo variant calls for SVs (per-sample and joint calling with *pbsv)* for the five HiFi read sets using the full or downsampled (5x) data sets.

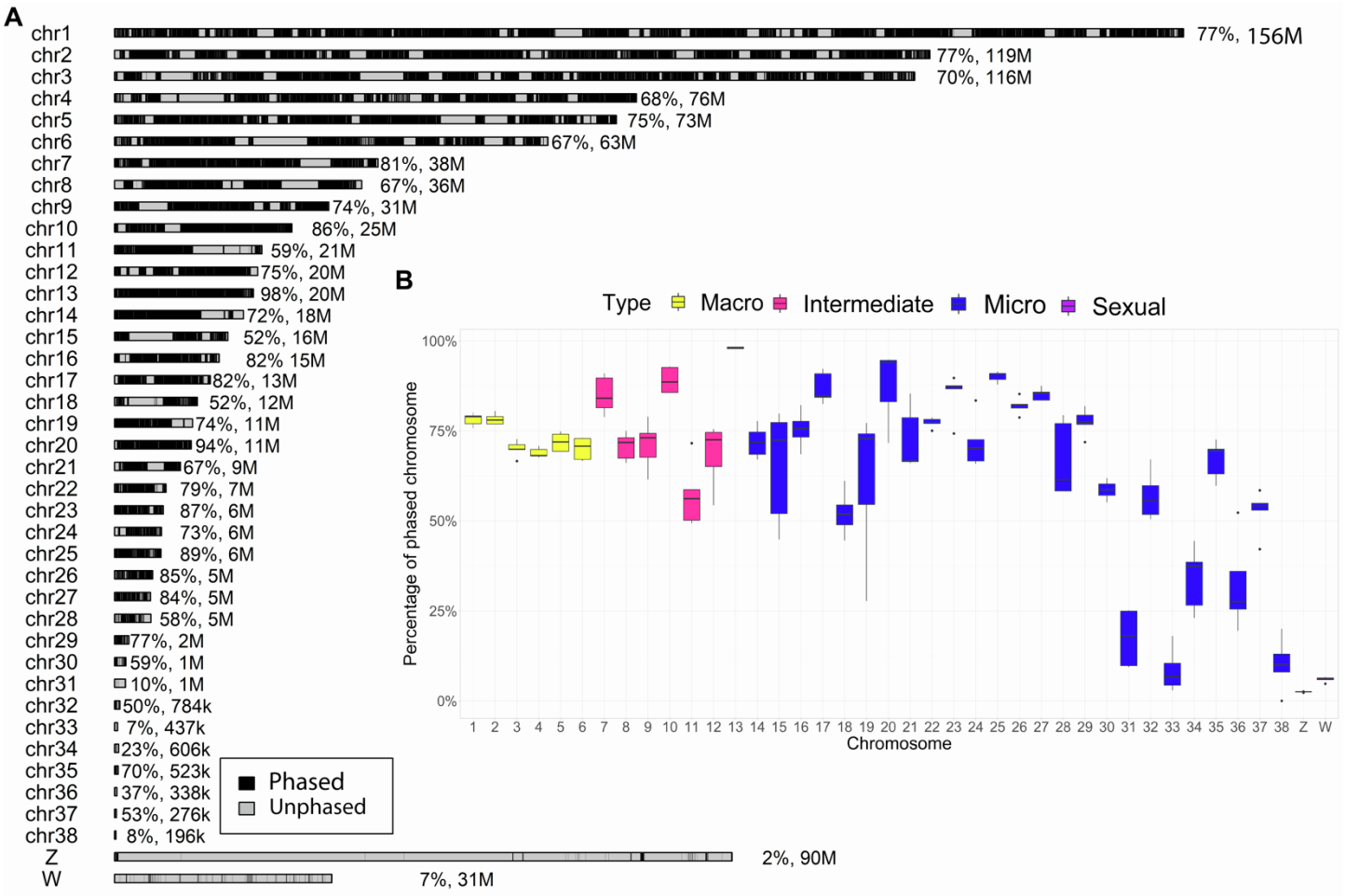


Figure S14. Haplotype phasing per chromosome with HiFi technology. Related to Data S1.
(A) Length of phased blocks (black) for sample A1 reported as example. Percentage of phasing and chromosome size is reported alongside each chromosome. **(B)** The boxplot shows the percentage of phased chromosomes, coloured by type, across all samples.

Supplementary Tables

Table S1 – Assembly statistics. Related to Figure 1 and Data S1. a) Genomescope2.0: Genome size, heterozygosity and repeat content prediction (p=2, k=31). b) Assembly statistics for the final VGP assembly; the VGP assembly before manual curation; the VGP assembly after the first round of curation (no genes recovery); Chelidonia assembly; Chelidonia assembly after haplotig purging; the short reads-based barn swallow assembly from Feng et al., 2020. c) *K*-mer completeness, QV and duplications computed with Merqury. d) BUSCO scores. e) Chelidonia haplotig purging results.

a) Genomescope2.0

Property	min	max	mean
Homozygous (aa) (%)	98.9601	98.9682	98.96415
Heterozygous (ab) (%)	1.03185	1.03988	1.035865
Genome Haploid Length (bp)	1,240,987,705	1,242,467,779	1,241,727,742
Genome Repeat Length (bp)	280,515,586	280,850,146	280,682,866
Genome Unique Length (bp)	960,472,119	961,617,634	961,044,877
Model Fit (%)	79.4413	98.3039	88.8726

b) asm_stats

Assembly	Level	Total bp	Number	Max length (bp)	N50 (bp)	N90 (bp)	NG50 (bp)	NG90 (bp)
bHirRus1	Scaffolds	1,105,955,550	617	156,035,725	76,187,387	12,073,725	73,257,097	-
	Contigs	1,080,421,138	1,719	18,822,688	2,794,774	410,547	2,317,408	-
	Gaps	25,534,411	1,102	859,144	67,235	25,050	-	-
bHirRus1 non-curated	Scaffolds	1,108,599,991	670	112,528,062	61,988,788	5,759,618	54,964,849	-
	Contigs	1,079,323,631	1,690	18,822,688	2,794,774	416,910	2,317,408	-
	Gaps	29,276,367	1,022	859,144	87,847	20,407	-	-
bHirRus1 before BUSCO gene recovery	Scaffolds	1,104,070,652	578	156,035,725	76,187,387	12,073,725	73,257,097	-
	Contigs	1,078,538,515	1,677	18,822,688	2,794,774	421,471	2,317,408	-
	Gaps	25,532,137	1,099	859,144	67,235	25,050	-	-
Chelidonia	Scaffolds	1,213,743,879	364	98,053,015	25,954,216	2,002,624	25,954,216	1,397,752
	Contigs	1,173,760,729	1.335	33,289,027	6,124,600	363.863	5,557,227	139.41
	Gaps	39,983,150	971	1,519,005	190,808	36,955	-	-
Chelidonia after haplotig purging	Scaffolds	1,158,593,414	301	97,796,368	30,885,980	2,436,503	25,929,213	734,124
	Contigs	1,118,668,438	995	33,289,027	6,435,469	591,306	5,538,306	34,521
	Gaps	39,925,031	747	1,721,071	281,095	50,550	-	-
Feng et al., 2020	Scaffolds	1,043,753,555	13,084	5,166,709	675,769	82,647	478,971	-
	Contigs	1,038,939,019	60,667	536,535	53,490	10,836	41,237	-
	Gaps	4,814,536	47,583	13,382	710	107	-	-

c) Merqury

Assembly	QV	Completeness
bHirRus1	43.7271	83.3386
Alternate assembly	41.1616	74.181
bHirRus1 + alternate	42.3478	93.8648
bHirRus1 before BUSCO gene recovery	44.0223	83.2454
bHirRus1 before BUSCO gene recovery + alternate	42.4594	93.7927
Chelidonia	33.9565	84.8622
Chelidonia after haplotig purging	34.2602	84.2029
Feng et al., 2020	24.3342	40.2206

False duplication content

Histogram	cutoff	1	2	3	4	>4	dup(>1)	all	dup%
bHirRus1 before BUSCO gene recovery	84	9.27E+08	3935932	253910	49034	27208	4266084	9.31E+08	0.458118
bHirRus1	84	9.28E+08	4218223	267726	54013	29560	4569522	9.32E+08	0.490137
c1	84	9.01E+08	36656566	707513	147460	62899	37574438	9.39E+08	4.00275
p1	84	9.22E+08	4258910	266200	53209	29325	4607644	9.27E+08	0.49727
Chelidonia	84	9.38E+08	11470565	870422	152316	74823	12568126	9.50E+08	1.32249
Chelidonia after haplotig purging	84	9.38E+08	4748015	323016	62305	39539	5172875	9.43E+08	0.54871

d) BUSCO

Assembly	Complete BUSCOs (C)	Complete and single-copy BUSCOs (S)	Complete and duplicated BUSCOs (D)	Fragmented BUSCOs (F)	Missing BUSCOs (M)	Total BUSCO groups searched	BUSCO stats
bHirRus1	3220	3192	28	37	97	3354	C:96.0%[S:95.2%, D:0.8%],F:1.1%,M:2.9%,n:3354
Alternate assembly	2926	2908	18	70	358	3354	C:87.2%[S:86.7%, D:0.5%],F:2.1%,M:10.7%,n:3354

bHirRus1 before BUSCO gene recovery	3181	3153	28	34	139	3354	C:94.8%[S:94.0%, D:0.8%],F:1.0%,M:4.2%,n:3354
Chelidonia	3215	3173	42	40	99	3354	C:95.9%[S:94.6%, D:1.3%],F:1.2%,M:2.9%,n:3354
Chelidonia after haplotig purging	3213	3185	28	40	101	3354	C:95.8%[S:95.0%, D:0.8%],F:1.2%,M:3.0%,n:3354

e) Chelidonia haplotig purging results

Type	n°	length (bp)
HIGHCOV	14	1661729
HAPLOTIG	268	27114574
OVLP	159	10465028
REPEAT	153	15814511
tot	594	55055842

Table S4 – Genome annotation. Related to STAR methods. a) RNAseq and Isoseq public data. b) Genomic features from annotation. The different features were obtained with *GenomicFeature* R package from the gff annotation file. Overlapping coordinates were merged for the calculations.

a) RNAseq and Isoseq public data

Strategy	Tissue	Instrument	Experiment accession n°	Run accession n*	Spots (Mb)	Bases (Gb)	Link
RNAseq	Ovary	Illumina NovaSeq 6000	SRX9927592	SRR13516425	56.9	17.2	https://www.ncbi.nlm.nih.gov/sra/SRX9927592[accn]
RNAseq	Brain	Illumina NovaSeq 6000	SRX9927591	SRR13516426	48.8	14.7	https://www.ncbi.nlm.nih.gov/sra/SRX9927591%5baccn
RNAseq	Muscle	Illumina NovaSeq 6000	SRX9927590	SRR13516427	54	16.3	https://www.ncbi.nlm.nih.gov/sra/12987859
RNAseq	Brain	Illumina HiSeq 2500	SRX7523274	SRR10853074	81	8.1	https://www.ncbi.nlm.nih.gov/sra/SRX7523274[accn
IsoSeq	Ovary	PacBio SMRT Sequel	SRX5956896	SRR9184408	0.29	0.097	https://www.ncbi.nlm.nih.gov/sra/SRX5956896%5baccn
IsoSeq	Brain	PacBio SMRT Sequel	SRX5956895	SRR9184409	0.24	0.077	https://www.ncbi.nlm.nih.gov/sra/SRX5956895%5baccn

b) Genomic features from annotation

Feature	total bp	% of the assembly	% of genes
Genes	573692145	51.9	-
Introns	524783002	47.5	91.5
Exons	55560589	5.0	9.7
CDS	28324328	2.6	4.9
5'UTRs	3952868	0.4	0.7
3'UTRs	19748013	1.8	3.4
Intergenic regions	532263405	48.1	92.8

Table S5 – Cactus alignment. Related to Figure 2C, S3A and STAR methods. a) Species used in Cactus alignment. The scientific name, the common name, the NCBI assembly name and accession number are shown for each species. Genome length and N50 were taken from NCBI. The table also report the number of masked bases and percentage of the genome that was masked, and also the bases aligned to the barn swallow and the percentage of the genome that aligned. The species excluded from the subsequent analysis are in grey. b) bHirRus1 Cactus alignment coverage per chromosome.

a) Cactus alignment species

Scientific name	Common name	Accession number	Tot bp	N50 (bp)	Masked bp	% masked	Aligned bp	% aligned
Molothrus ater	Brown-headed cowbird	GCF_012460135.1	1,087,312,585	52,124,711	269,471,518	24.78	881,341,738	79.69
Motacilla alba	White wagtail	GCF_015832195.1	1,072,670,728	72,386,170	272,868,451	25.44	880,594,475	79.62
Camarhynchus parvulus	Small tree finch	GCF_901933205.1	1,051,609,828	70,356,807	261,698,129	24.89	874,775,347	79.09
Passer domesticus	House sparrow	GCA_001700915.1	1,042,720,703	6,373,860	228,033,610	21.87	871,458,883	78.79
Lonchura striata domestica	Bengalese finch	GCF_005870125.1	1,060,269,806	71,975,342	245,415,318	23.15	863,868,784	78.11
Taenopygia guttata	Zebra finch	GCF_008822105.2	1,068,988,109	70,879,221	256,101,562	23.96	860,556,598	77.81
Gallus gallus	Chicken	GCF_016699485.2	1,053,332,251	90,861,225	264,817,877	25.14	662,165,156	59.87
Parus major	Great tit	GCF_001522545.3	1,020,310,769	71,365,269	225,465,405	22.10	27	2.65
Ficedula albicollis	Collared flycatcher	GCF_000247815.1	1,118,343,587	6,542,656	296,963,156	26.55	80	7.15

b) Cactus alignment coverage per chromosome

Chr	Size (bp)	0 genomes	1 genome	2 genomes	3 genomes	4 genomes	5 genomes	6 genomes	7 genomes	% aligned	% not aligned
1	156035725	22532775	94404	109679	274008	1559855	5332627	35242269	90890108	85.56	14.44
2	119023421	17120388	86564	114723	298724	1425231	4657566	27027733	68292492	85.62	14.38
3	116801625	16367836	70252	90631	191620	1086574	3781877	25048151	70164684	85.99	14.01
Z	90132487	1759600	12899	16001	44761	200355	681347	2674231	4841346	98.05	1.95
4	76187387	12969937	56265	77688	189633	875693	3194956	15721209	43102006	82.98	17.02
5	73257097	10014716	78449	70938	166941	810028	2720606	16710628	42684791	86.33	13.67
6	63258489	7934165	53327	58950	143002	757292	2322007	14350140	37639606	87.46	12.54
7	38459648	3870322	36235	43369	106206	491506	1440963	8797534	23673513	89.94	10.06
8	36085389	4473612	30249	44714	115394	465077	1451433	8003464	21501446	87.60	12.40
W	31704074	26864875	24743	95090	234938	166864	303379	1344969	2669216	15.26	84.74
9	31262510	2994007	27026	38765	107071	431152	1252595	7358268	19053626	90.42	9.58
10	25880253	2829335	26193	42310	102602	435300	1239054	6419206	14786253	89.07	10.93
11	21491857	2491859	19301	25876	63611	261302	795016	4764344	13070548	88.41	11.59
12	20890524	2788498	32787	30390	76214	263131	789694	4803906	12105904	86.65	13.35
13	20272128	2023772	16771	21061	53703	219746	768595	4427105	12741375	90.02	9.98
14	18810845	2452724	24623	45298	73997	283376	818648	4282907	10829272	86.96	13.04
15	16541138	1579808	24375	21016	60815	245029	747715	4346476	9515904	90.45	9.55
16	15277844	2336924	16522	26656	63252	223732	646004	3526897	8437857	84.70	15.30
17	13985943	1721483	16317	21378	53879	210481	650277	3460215	7851913	87.69	12.31
18	12073725	2267587	27237	34654	99275	301281	772153	2840829	5730709	81.22	18.78
19	11382101	1328753	17196	24082	51016	205288	547100	2798493	6410173	88.33	11.67
20	11194341	1345733	13281	17046	42646	158865	464075	2592367	6560328	87.98	12.02
21	9617204	1179314	9355	11652	23613	100480	353390	2300716	5638684	87.74	12.26
22	7507825	1531678	17867	30822	79264	203219	453555	1961827	3229593	79.60	20.40
23	7098401	1076232	15562	17156	34376	131589	392658	1913710	3517118	84.84	15.16
24	6843954	1590607	17142	25826	64027	196198	498113	1893733	2558308	76.76	23.24
25	6778862	1353172	19566	26353	66240	211551	515989	1911842	2674149	80.04	19.96
26	5553549	1252173	18536	20175	54283	163204	463395	1571580	2010203	77.45	22.55
27	5236451	1363306	16945	30470	96822	236583	484255	1337925	1670145	73.97	26.03
28	5297670	2172676	28661	34797	107405	214801	409558	976611	1353161	58.99	41.01
29	2102120	636352	14911	16723	43901	100617	238219	494811	556586	69.73	30.27
30	1648998	626826	11081	14520	40326	94972	174771	312494	374008	61.99	38.01
31	1590086	1264350	10791	4317	15874	23373	39040	107858	124483	20.49	79.51
32	784579	716285	12234	1494	12182	12594	12903	15680	1207	8.70	91.30
33	437724	407203	729	2599	2293	2507	6662	11940	3791	6.97	93.03
34	606149	583775	147	498	3153	1192	1490	11481	4413	3.69	96.31
35	523230	154497	1361	3700	4307	17733	44006	112568	185058	70.47	29.53
36	338027	178933	5768	12342	32148	51294	30589	18316	8637	47.07	52.93
37	276370	201909	665	4820	7982	20097	11709	14798	14390	26.94	73.06
38	220485	192908	1194	1565	5353	10409	7359	508	1189	12.51	87.49
39	65965	54934	129	1010	1580	5189	2569	25	529	16.72	83.28

Table S6 – Selection analysis. Related to Figure 2C, S3 and STAR methods. a) Total conserved and accelerated elements and bases computed with PhyloP and PhastCons and corrected with a 5% FDR threshold. Points above the horizontal line are significant also according to Bonferroni correction. No conserved PhyloP sites were significant after Bonferroni correction. The chromosomal fraction (Chr fraction column) was calculated as the percentage of the bHirRus1 chromosomal bases (1,082,536,200 bp) covered by conserved or accelerated bases. b) Number and percentage of FDR corrected conserved and accelerated bases that fall into different genomic features (extracted with GenomicFeatures R package). c) Base pair number of genomic features in the chromosomes and percentage of genomic features bases that are covered by conserved or accelerated elements.

a)	**PhyloP acc. (FDR)**	**PhyloP acc. (FDR + Bonferroni)**	**PhyloP cons. (FDR)**	**PhastCons**
total (bp)	10364130	63710	29303310	132672359
n° elements	1036413	6371	2930331	2961683
Chr fraction (%)	0.96	0.01	2.71	12.26
b)				
Genic sites (bp)	5343488	32488	18317624	80706942
Genic sites (%)	51.56	50.99	62.51	60.83
CDS sites (bp)	93756	892	5011640	18053504
CDS sites (%)	0.90	1.40	17.10	13.61
5' UTRs sites (bp)	65325	673	226695	911705
5' UTRs sites (%)	0.63	1.06	0.77	0.69
3' UTRs sites (bp)	142070	1182	1674976	5520325
3' UTRs sites (%)	1.37	1.86	5.72	4.16
Intronic sites (bp)	5010136	29421	11340733	55894466
Intronic sites (%)	48.34	46.18	38.70	42.13
Intergenic sites (bp)	5020642	31222	10985686	51965405
Intergenic sites (%)	48.44	49.01	37.49	39.17
c)				
Genes (bp)	567449897	567449897	567449897	567449897
Genes under selection (%)	0.02	0.00	0.88	3.18
CDS (bp)	27338408	27338408	27338408	27338408
CDS under selection (%)	0.34	0.00	18.33	66.04
5' UTRs (bp)	3857953	3857953	3857953	3857953
5' UTRs under selection (%)	1.69	0.02	5.88	23.63
3' UTRs (bp)	19510415	19510415	19510415	19510415
3' UTRs under selection (%)	0.73	0.01	8.59	28.29
Introns (bp)	519872436	519872436	519872436	519872436
Introns under selection (%)	0.96	0.01	2.18	10.75
Intergenes (bp)	515086303	515086303	515086303	515086303
Intergenes under selection (%)	0.97	0.01	2.13	10.09

Table S8 - Gene ontology analysis on PhyloP accelerated genes. Relates to Figure S3E. The top 5% (606) genes with more accelerated bases overlapped with the entire gene sequence ("LOC" genes excluded) were tested. The first 100 lines are showed.

GO term	p.geomean	stat.mean	p.val	q.val	set.size	exp1	GO term name
GO:0005840	0.0027	2.8106	0.0027	0.9111	131	0.0027	ribosome
GO:0003735	0.0031	2.7743	0.0031	0.9111	105	0.0031	structural constituent of ribosome
GO:0006614	0.0054	2.6016	0.0054	0.9111	63	0.0054	SRP-dependent cotranslational protein targeting to membrane
GO:0000184	0.0075	2.4644	0.0075	0.9111	85	0.0075	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
GO:0002181	0.0082	2.4474	0.0082	0.9111	59	0.0082	cytoplasmic translation
GO:0032729	0.0088	2.4620	0.0088	0.9111	35	0.0088	positive regulation of interferon-gamma production
GO:0006397	0.0128	2.2382	0.0128	0.9111	277	0.0128	mRNA processing
GO:0022626	0.0131	2.2661	0.0131	0.9111	56	0.0131	cytosolic ribosome
GO:0008380	0.0135	2.2212	0.0135	0.9111	223	0.0135	RNA splicing
GO:0002250	0.0139	2.2252	0.0139	0.9111	87	0.0139	adaptive immune response
GO:0006413	0.0149	2.1914	0.0149	0.9111	105	0.0149	translational initiation
GO:0042104	0.0168	2.3365	0.0168	0.9111	13	0.0168	positive regulation of activated T cell proliferation
GO:0019083	0.0173	2.1370	0.0173	0.9111	78	0.0173	viral transcription
GO:0032418	0.0180	2.1940	0.0180	0.9111	24	0.0180	lysosome localization
GO:0022625	0.0216	2.0696	0.0216	0.9111	40	0.0216	cytosolic large ribosomal subunit
GO:0001772	0.0238	2.0453	0.0238	0.9111	29	0.0238	immunological synapse
GO:0050821	0.0242	1.9847	0.0242	0.9111	139	0.0242	protein stabilization
GO:0006412	0.0294	1.8952	0.0294	0.9111	222	0.0294	translation
GO:0008047	0.0349	1.8508	0.0349	0.9111	37	0.0349	enzyme activator activity
GO:0005681	0.0370	1.7959	0.0370	0.9111	120	0.0370	spliceosomal complex
GO:0000398	0.0399	1.7570	0.0399	0.9111	179	0.0399	mRNA splicing, via spliceosome
GO:0032740	0.0402	1.8593	0.0402	0.9111	14	0.0402	positive regulation of interleukin-17 production
GO:0015935	0.0402	1.8593	0.0402	0.9111	14	0.0402	small ribosomal subunit
GO:0006364	0.0409	1.7469	0.0409	0.9111	161	0.0409	rRNA processing
GO:0044183	0.0435	1.7564	0.0435	0.9111	28	0.0435	protein folding chaperone
GO:0070125	0.0456	1.7074	0.0456	0.9111	59	0.0456	mitochondrial translational elongation
GO:0048704	0.0465	1.7044	0.0465	0.9111	44	0.0465	embryonic skeletal system morphogenesis
GO:0007040	0.0499	1.6818	0.0499	0.9111	31	0.0499	lysosome organization
GO:0070126	0.0504	1.6568	0.0504	0.9111	62	0.0504	mitochondrial translational termination
GO:0005249	0.0540	1.6269	0.0540	0.9111	48	0.0540	voltage-gated potassium channel activity
GO:1904724	0.0544	1.6348	0.0544	0.9111	33	0.0544	tertiary granule lumen
GO:0009952	0.0554	1.6052	0.0554	0.9111	82	0.0554	anterior/posterior pattern specification
GO:0043236	0.0568	1.6364	0.0568	0.9111	20	0.0568	laminin binding
GO:0051015	0.0579	1.5775	0.0579	0.9111	158	0.0579	actin filament binding
GO:0002230	0.0597	1.6059	0.0597	0.9111	21	0.0597	positive regulation of defense response to virus by host
GO:0002376	0.0627	1.5353	0.0627	0.9111	264	0.0627	immune system process
GO:0006369	0.0658	1.5481	0.0658	0.9111	23	0.0658	termination of RNA polymerase II transcription
GO:0005685	0.0662	1.6184	0.0662	0.9111	10	0.0662	U1 snRNP
GO:0006898	0.0666	1.5122	0.0666	0.9111	71	0.0666	receptor-mediated endocytosis
GO:1990904	0.0682	1.4958	0.0682	0.9111	108	0.0682	ribonucleoprotein complex
GO:0018279	0.0689	1.5206	0.0689	0.9111	24	0.0689	protein N-linked glycosylation via asparagine
GO:0022627	0.0689	1.5206	0.0689	0.9111	24	0.0689	cytosolic small ribosomal subunit
GO:0003729	0.0692	1.4863	0.0692	0.9111	147	0.0692	mRNA binding
GO:0031083	0.0705	1.5676	0.0705	0.9111	11	0.0705	BLOC-1 complex
GO:0001833	0.0705	1.5676	0.0705	0.9111	11	0.0705	inner cell mass cell proliferation
GO:0097623	0.0705	1.5676	0.0705	0.9111	11	0.0705	potassium ion export across plasma membrane
GO:0010614	0.0705	1.5676	0.0705	0.9111	11	0.0705	negative regulation of cardiac muscle hypertrophy
GO:0008250	0.0705	1.5676	0.0705	0.9111	11	0.0705	oligosaccharyltransferase complex
GO:0007625	0.0705	1.5676	0.0705	0.9111	11	0.0705	grooming behavior
GO:0045667	0.0705	1.5676	0.0705	0.9111	11	0.0705	regulation of osteoblast differentiation
GO:0071005	0.0717	1.4827	0.0717	0.9111	40	0.0717	U2-type precatalytic spliceosome
GO:0032735	0.0721	1.4939	0.0721	0.9111	25	0.0721	positive regulation of interleukin-12 production
GO:1904813	0.0745	1.4504	0.0745	0.9111	93	0.0745	ficolin-1-rich granule lumen
GO:0099524	0.0747	1.5222	0.0747	0.9111	12	0.0747	postsynaptic cytosol
GO:0043202	0.0768	1.4357	0.0768	0.9111	76	0.0768	lysosomal lumen
GO:0007166	0.0782	1.4211	0.0782	0.9111	153	0.0782	cell surface receptor signaling pathway
GO:0006506	0.0786	1.4425	0.0786	0.9111	27	0.0786	GPI anchor biosynthetic process
GO:0071007	0.0786	1.4425	0.0786	0.9111	27	0.0786	U2-type catalytic step 2 spliceosome
GO:0086005	0.0790	1.4809	0.0790	0.9111	13	0.0790	ventricular cardiac muscle cell action potential
GO:1900029	0.0790	1.4809	0.0790	0.9111	13	0.0790	positive regulation of ruffle assembly
GO:0071157	0.0790	1.4809	0.0790	0.9111	13	0.0790	negative regulation of cell cycle arrest
GO:0030425	0.0803	1.4048	0.0803	0.9111	368	0.0803	dendrite
GO:0043025	0.0829	1.3876	0.0829	0.9111	326	0.0829	neuronal cell body
GO:0030030	0.0830	1.3890	0.0830	0.9111	156	0.0830	cell projection organization

GO:0101031	0.0833	1.4427	0.0833	0.9111	14	0.0833	chaperone complex
GO:0060307	0.0833	1.4427	0.0833	0.9111	14	0.0833	regulation of ventricular cardiac muscle cell membrane repolarization
GO:0048489	0.0833	1.4427	0.0833	0.9111	14	0.0833	synaptic vesicle transport
GO:0097067	0.0833	1.4427	0.0833	0.9111	14	0.0833	cellular response to thyroid hormone stimulus
GO:0001736	0.0833	1.4427	0.0833	0.9111	14	0.0833	establishment of planar polarity
GO:0048026	0.0877	1.4067	0.0877	0.9111	15	0.0877	positive regulation of mRNA splicing, via spliceosome
GO:1901379	0.0877	1.4067	0.0877	0.9111	15	0.0877	regulation of potassium ion transmembrane transport
GO:0050776	0.0884	1.3637	0.0884	0.9111	46	0.0884	regulation of immune response
GO:1901224	0.0944	1.3257	0.0944	0.9111	48	0.0944	positive regulation of NIK/NF-kappaB signaling
GO:0008076	0.0957	1.3143	0.0957	0.9111	66	0.0957	voltage-gated potassium channel complex
GO:0043198	0.0961	1.3229	0.0961	0.9111	32	0.0961	dendritic shaft
GO:0035145	0.0967	1.3403	0.0967	0.9111	17	0.0967	exon-exon junction complex
GO:0019882	0.0967	1.3403	0.0967	0.9111	17	0.0967	antigen processing and presentation
GO:0044325	0.0969	1.3039	0.0969	0.9111	104	0.0969	ion channel binding
GO:0007052	0.0992	1.2910	0.0992	0.9111	105	0.0992	mitotic spindle organization
GO:0015459	0.0997	1.3003	0.0997	0.9111	33	0.0997	potassium channel regulator activity
GO:0009925	0.0997	1.3003	0.0997	0.9111	33	0.0997	basal plasma membrane
GO:0042102	0.0997	1.3003	0.0997	0.9111	33	0.0997	positive regulation of T cell proliferation
GO:0051603	0.0997	1.3003	0.0997	0.9111	33	0.0997	proteolysis involved in cellular protein catabolic process
GO:0016032	0.1006	1.2792	0.1006	0.9111	409	0.1006	viral process
GO:0034599	0.1012	1.2822	0.1012	0.9111	68	0.1012	cellular response to oxidative stress
GO:0008202	0.1012	1.2822	0.1012	0.9111	68	0.1012	steroid metabolic process
GO:0045672	0.1013	1.3091	0.1013	0.9111	18	0.1013	positive regulation of osteoclast differentiation
GO:0044306	0.1013	1.3091	0.1013	0.9111	18	0.1013	neuron projection terminus
GO:0005244	0.1037	1.2652	0.1037	0.9111	107	0.1037	voltage-gated ion channel activity
GO:0008198	0.1060	1.2791	0.1060	0.9111	19	0.1060	ferrous iron binding
GO:0043113	0.1060	1.2791	0.1060	0.9111	19	0.1060	receptor clustering
GO:0031369	0.1060	1.2791	0.1060	0.9111	19	0.1060	translation initiation factor binding
GO:0048306	0.1070	1.2520	0.1070	0.9111	52	0.1070	calcium-dependent protein binding
GO:0032438	0.1108	1.2500	0.1108	0.9111	20	0.1108	melanosome organization
GO:0051602	0.1108	1.2500	0.1108	0.9111	20	0.1108	response to electrical stimulus
GO:0005743	0.1122	1.2164	0.1122	0.9111	304	0.1122	mitochondrial inner membrane
GO:0001669	0.1136	1.2161	0.1136	0.9111	54	0.1136	acrosomal vesicle
GO:0042277	0.1136	1.2161	0.1136	0.9111	54	0.1136	peptide binding
GO:0004896	0.1150	1.2129	0.1150	0.9111	37	0.1150	cytokine receptor activity
GO:0034765	0.1154	1.2018	0.1154	0.9111	112	0.1154	regulation of ion transmembrane transport

Table S10 - Gene ontology analysis on PhyloP conserved genes. Related to Figure S3F. The top 5% (606) genes with more accelerated bases overlapped with CDS ("LOC" genes excluded) were tested. The first 100 lines are showed.

GO term	p.geomean	stat.mean	p.val	q.val	set.size	exp1	GO term name
GO:1990837	1.24E-08	5.646982295	1.24E-08	4.22E-05	394	1.24E-08	sequence-specific double-stranded DNA binding
GO:0045892	3.61E-08	5.448719491	3.61E-08	6.12E-05	407	3.61E-08	negative regulation of transcription, DNA-templated
GO:0003700	6.63E-08	5.325497347	6.63E-08	7.50E-05	469	6.63E-08	DNA-binding transcription factor activity
GO:0043565	5.22E-07	4.939109527	5.22E-07	0.000442879	343	5.22E-07	sequence-specific DNA binding
GO:0001228	1.66E-06	4.701596018	1.66E-06	0.001123946	324	1.66E-06	DNA-binding transcription activator activity, RNA polymerase II-specific
GO:0008134	1.73E-05	4.191052732	1.73E-05	0.009768698	240	1.73E-05	transcription factor binding
GO:0007399	2.81E-05	4.051681629	2.81E-05	0.01360715	440	2.81E-05	nervous system development
GO:0005667	6.00E-05	3.906384172	6.00E-05	0.025428411	166	6.00E-05	transcription regulator complex
GO:0000977	0.000126182	3.695609073	0.000126182	0.047570723	229	0.000126182	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0000165	0.000335715	3.429373373	0.000335715	0.113908255	230	0.000335715	MAPK cascade
GO:0000987	0.000484272	3.417081778	0.000484272	0.14364186	61	0.000484272	cis-regulatory region sequence-specific DNA binding
GO:0009887	0.000508017	3.338363757	0.000508017	0.14364186	122	0.000508017	animal organ morphogenesis
GO:0009952	0.000560403	3.338674072	0.000560403	0.146265234	82	0.000560403	anterior/posterior pattern specification
GO:0070936	0.000673827	3.347672489	0.000673827	0.163306678	48	0.000673827	protein K48-linked ubiquitination
GO:0001227	0.000942588	3.13560358	0.000942588	0.213213356	183	0.000942588	DNA-binding transcription repressor activity, RNA polymerase II-specific
GO:0061631	0.001020932	3.382977443	0.001020932	0.216501408	23	0.001020932	ubiquitin conjugating enzyme activity
GO:0008284	0.001413664	2.997248451	0.001413664	0.282150649	370	0.001413664	positive regulation of cell population proliferation
GO:0006511	0.001586101	2.969629434	0.001586101	0.298980004	223	0.001586101	ubiquitin-dependent protein catabolic process
GO:0098978	0.0018269	2.92000323	0.0018269	0.326245811	295	0.0018269	glutamatergic synapse
GO:0031625	0.001944592	2.905542961	0.001944592	0.329422708	216	0.001944592	ubiquitin protein ligase binding
GO:1900740	0.002038867	3.244409302	0.002038867	0.329422708	16	0.002038867	positive regulation of protein insertion into mitochondrial membrane involved in apoptotic signaling pathway
GO:0010468	0.002367351	2.841389276	0.002367351	0.36511014	222	0.002367351	regulation of gene expression
GO:0003682	0.002736452	2.787901134	0.002736452	0.401281403	338	0.002736452	chromatin binding
GO:0019003	0.002838418	2.837093262	0.002838418	0.401281403	59	0.002838418	GDP binding
GO:0010628	0.002995891	2.757488391	0.002995891	0.406602358	356	0.002995891	positive regulation of gene expression
GO:0003151	0.003492259	2.79302391	0.003492259	0.45573976	43	0.003492259	outflow tract morphogenesis
GO:0048704	0.003750725	2.764175965	0.003750725	0.471341059	44	0.003750725	embryonic skeletal system morphogenesis
GO:0000976	0.004417561	2.634543109	0.004417561	0.535313765	188	0.004417561	transcription regulatory region sequence-specific DNA binding
GO:0030182	0.004580254	2.626522148	0.004580254	0.535889695	150	0.004580254	neuron differentiation
GO:0010629	0.004955134	2.592999359	0.004955134	0.560425688	205	0.004955134	negative regulation of gene expression
GO:0051549	0.005752803	2.931801726	0.005752803	0.59802306	11	0.005752803	positive regulation of keratinocyte migration
GO:0008083	0.006198829	2.52690836	0.006198829	0.59802306	108	0.006198829	growth factor activity
GO:0001654	0.006369607	2.593571039	0.006369607	0.59802306	33	0.006369607	eye development
GO:0030334	0.006495688	2.523690533	0.006495688	0.59802306	74	0.006495688	regulation of cell migration
GO:0003729	0.006528657	2.50038294	0.006528657	0.59802306	147	0.006528657	mRNA binding
GO:0021983	0.006628593	2.617049945	0.006628593	0.59802306	25	0.006628593	pituitary gland development
GO:0003925	0.006628593	2.617049945	0.006628593	0.59802306	25	0.006628593	obsolete small monomeric GTPase activity
GO:0043484	0.006874707	2.559594307	0.006874707	0.59802306	34	0.006874707	regulation of RNA splicing
GO:0010975	0.006874707	2.559594307	0.006874707	0.59802306	34	0.006874707	regulation of neuron projection development
GO:0003924	0.007084287	2.465237564	0.007084287	0.59802306	203	0.007084287	GTPase activity
GO:0003690	0.007355656	2.466071095	0.007355656	0.59802306	100	0.007355656	double-stranded DNA binding
GO:0001658	0.007402584	2.526790881	0.007402584	0.59802306	35	0.007402584	branching involved in ureteric bud morphogenesis
GO:0030324	0.009479826	2.376130804	0.009479826	0.740119143	82	0.009479826	lung development

GO:0090575	0.009840225	2.374329876	0.009840225	0.740119143	60	0.009840225	RNA polymerase II transcription regulator complex
GO:0001656	0.010067244	2.417325816	0.010067244	0.740119143	30	0.010067244	metanephros development
GO:0038095	0.010362212	2.340774231	0.010362212	0.740119143	84	0.010362212	Fc-epsilon receptor signaling pathway
GO:0000209	0.010667119	2.312241225	0.010667119	0.740119143	203	0.010667119	protein polyubiquitination
GO:0007223	0.010845416	2.382500879	0.010845416	0.740119143	31	0.010845416	Wnt signaling pathway, calcium modulating pathway
GO:0048856	0.010851519	2.343853392	0.010851519	0.740119143	51	0.010851519	anatomical structure development
GO:0008543	0.01090656	2.332048968	0.01090656	0.740119143	62	0.01090656	fibroblast growth factor receptor signaling pathway
GO:0004842	0.011990904	2.2676406	0.011990904	0.797747797	193	0.011990904	ubiquitin-protein transferase activity
GO:0006417	0.012317927	2.265337942	0.012317927	0.802021374	113	0.012317927	regulation of translation
GO:0006325	0.013018603	2.233432465	0.013018603	0.802021374	239	0.013018603	chromatin organization
GO:0021762	0.013377186	2.285110766	0.013377186	0.802021374	34	0.013377186	substantia nigra development
GO:0003730	0.013467378	2.252202474	0.013467378	0.802021374	55	0.013467378	mRNA 3'-UTR binding
GO:0060076	0.013711241	2.305151624	0.013711241	0.802021374	25	0.013711241	excitatory synapse
GO:0048536	0.013711241	2.305151624	0.013711241	0.802021374	25	0.013711241	spleen development
GO:0043488	0.013933744	2.221069906	0.013933744	0.802021374	91	0.013933744	regulation of mRNA stability
GO:0031954	0.01412248	2.350517468	0.01412248	0.802021374	17	0.01412248	positive regulation of protein autophosphorylation
GO:0042752	0.014182518	2.230090501	0.014182518	0.802021374	56	0.014182518	regulation of circadian rhythm
GO:0001501	0.014782176	2.191746247	0.014782176	0.822228253	118	0.014782176	skeletal system development
GO:0007420	0.015682963	2.160259043	0.015682963	0.822530709	217	0.015682963	brain development
GO:0030900	0.015689729	2.186716873	0.015689729	0.822530709	58	0.015689729	forebrain development
GO:0030326	0.015751253	2.194599639	0.015751253	0.822530709	47	0.015751253	embryonic limb morphogenesis
GO:0048935	0.015999713	2.396877951	0.015999713	0.822530709	11	0.015999713	peripheral nervous system neuron development
GO:0001823	0.015999713	2.396877951	0.015999713	0.822530709	11	0.015999713	mesonephros development
GO:0010494	0.016482794	2.165428424	0.016482794	0.834718206	59	0.016482794	cytoplasmic stress granule
GO:0009653	0.016987461	2.138951967	0.016987461	0.838640287	96	0.016987461	anatomical structure morphogenesis
GO:0060425	0.017124574	2.240496875	0.017124574	0.838640287	19	0.017124574	lung morphogenesis
GO:0016055	0.017545139	2.11650851	0.017545139	0.838640287	178	0.017545139	Wnt signaling pathway
GO:0050679	0.017548913	2.14667746	0.017548913	0.838640287	49	0.017548913	positive regulation of epithelial cell proliferation
GO:0007254	0.018305307	2.140394571	0.018305307	0.852251414	39	0.018305307	JNK cascade
GO:0003712	0.01833609	2.106804594	0.01833609	0.852251414	98	0.01833609	transcription coregulator activity
GO:0042056	0.018692344	2.191918015	0.018692344	0.857069234	20	0.018692344	chemoattractant activity
GO:0005525	0.019347222	2.073323422	0.019347222	0.866307018	255	0.019347222	GTP binding
GO:0000381	0.01940446	2.113455478	0.01940446	0.866307018	40	0.01940446	regulation of alternative mRNA splicing, via spliceosome
GO:0051403	0.020309554	2.146599881	0.020309554	0.877360576	21	0.020309554	stress-activated MAPK cascade
GO:0019904	0.0213354	2.034826342	0.0213354	0.877360576	185	0.0213354	protein domain specific binding
GO:0106310	0.021619568	2.026508624	0.021619568	0.877360576	275	0.021619568	protein serine kinase activity
GO:0106311	0.021619568	2.026508624	0.021619568	0.877360576	275	0.021619568	protein threonine kinase activity
GO:0007409	0.021651726	2.041571986	0.021651726	0.877360576	77	0.021651726	axonogenesis
GO:0017053	0.021722346	2.061188641	0.021722346	0.877360576	42	0.021722346	transcription repressor complex
GO:0048538	0.021722346	2.061188641	0.021722346	0.877360576	42	0.021722346	thymus development
GO:0021537	0.021979266	2.103997674	0.021979266	0.877360576	22	0.021979266	telencephalon development
GO:0046329	0.021979266	2.103997674	0.021979266	0.877360576	22	0.021979266	negative regulation of JNK cascade
GO:0003007	0.022942506	2.035786411	0.022942506	0.880049865	43	0.022942506	heart morphogenesis
GO:0045778	0.023084126	2.14526608	0.023084126	0.880049865	14	0.023084126	positive regulation of ossification
GO:0006376	0.023084126	2.14526608	0.023084126	0.880049865	14	0.023084126	mRNA splice site selection
GO:0046827	0.023084126	2.14526608	0.023084126	0.880049865	14	0.023084126	positive regulation of protein export from nucleus
GO:0043161	0.023727107	1.99185778	0.023727107	0.884990003	146	0.023727107	proteasome-mediated ubiquitin-dependent protein catabolic process
GO:0045665	0.023735364	2.011027281	0.023735364	0.884990003	55	0.023735364	negative regulation of neuron differentiation
GO:0003714	0.024405902	1.980730889	0.024405902	0.88660381	133	0.024405902	transcription corepressor activity
GO:0001934	0.025186892	1.96593398	0.025186892	0.88660381	148	0.025186892	positive regulation of protein phosphorylation
GO:1990090	0.025347677	2.002810835	0.025347677	0.88660381	34	0.025347677	cellular response to nerve growth factor stimulus
GO:0048557	0.025466995	2.082873982	0.025466995	0.88660381	15	0.025466995	embryonic digestive tract morphogenesis
GO:0048663	0.0254865	2.02533112	0.0254865	0.88660381	24	0.0254865	neuron fate commitment
GO:0032436	0.026076936	1.968076542	0.026076936	0.88660381	57	0.026076936	positive regulation of proteasomal ubiquitin-dependent protein catabolic process
GO:1990904	0.026269246	1.951343376	0.026269246	0.88660381	108	0.026269246	ribonucleoprotein complex

GO:0048156	0.026883645	1.974110394	0.026883645	0.88660381	35	0.026883645	tau protein binding
GO:0006605	0.026883645	1.974110394	0.026883645	0.88660381	35	0.026883645	protein targeting

Table S11 – Barn swallow individuals sequenced with HiFi technology. **Related to Figure 3, Figure 5, Figure S5, S6 and S7, STAR methods and Data S1.** a) Samples information. b) HiFi sequencing statistics. c) HiFi assembly statistics. d) Genomescope2.0 predictions based on HiFi raw data. e) Custom purge_dups cutoffs for HiFi. The k-mer coverage (kcov) was computed with genomescope2.0. f) Haplotig purging on HiFi assemblies. g) Alignment metrics from pbmm2 output for all HiFi samples (ds1, see Table S12).

a) Samples					
Sample name	**2**	**3**	**4**	**A1**	**A2**
Location	Parco Adda (CR)	Brescia	Oleggio (NO)	Matera	Matera
Latitude	45°26′N	45°22′N	45°34′N	40°40′N	40°40′N
Longitude	9°30′E	10°15'E	8°39'E	16°36′E	16°36′E
Sex	F	F	F	F	F
Sequencing coverage	25x	19.5x	20x	33x	15x
b) Sequencing statistics					
>= Q20 Reads	2,493,595	940,660	1,856,659	1,530,055	1,430,811
>= Q20 Yield (bp)	36,551,797,585	17,101,319,656	27,772,551,635	21,646,429,144	22,740,719,734
>= Q20 Read Lenght (mean, bp)	14657.5	18,180	14,928	14,147	16,136
>= Q20 Read Quality (median)	Q32	Q29	Q32	Q32	Q31
Sequencing coverage	33	15	25	19	20
>= Q20 Reads	2,493,595	940,660	1,856,659	1,530,055	1,430,811

c) HiFi assemblies statistics

	Primary before purging		**Primary after purging**			**Alternate before purging**		**Alternate after purging**		
Sample 2	**Contigs**	**Gaps**	**Contigs**	**Gaps**	**Purged (bp)**	**Contigs**	**Gaps**	**Contigs**	**Gaps**	**Purged (bp)**
Total bp	1,307,545,121	0	1,163,058,884	1,288	144,486,237	1,106,170,783	0	1,011,189,907	138	94,980,876
Number	1,959	0	1,106	56	-	4,940	0	2,840	6	-
Max length (bp)	40,208,769	0	40,208,769	23	-	10,650,792	0	10,650,792	23	-
N50 (bp)	5,371,840	0	7,140,827	23	-	1,099,778	0	1,225,108	23	-
N90 (bp)	295,307	0	564,576	23	-	89,329	0	198,244	23	-
NG50 (bp)	8,189,401	0	8,189,401	-	-	1,165,674	0	1,114,799	-	-
NG90 (bp)	1,339,570	0	1,156,149	-	-	166,867	0	51,878	-	-
Sample 3										
Total bp	1,273,430,777	0	1,135,657,041	1,081	137,773,736	1,072,637,793	0	991,522,024	138	81,115,769
Number	2,368	0	1,434	47	-	6,324	0	3,941	6	-
Max length (bp)	38,082,319	0	38,082,319	23	-	4,654,537	0	4,654,537	23	-
N50 (bp)	3,477,328	0	4,230,173	23	-	579,024	0	636,015	23	-
N90 (bp)	216,449	0	405,189	23	-	66,297	0	117,718	23	-
NG50 (bp)	4,473,894	0	4,473,894	-	-	567,346	0	585,425	-	-
NG90 (bp)	887,578	0	753,224	-	-	55,473	0	43,696	-	-
Sample 4										
Total bp	1,301,377,510	0	1,139,278,982	1,219	162,098,528	1,072,102,039	0	984,724,034	69	87,378,005
Number	2,199	0	1,267	53	-	5,743	0	3,549	3	-
Max length (bp)	21,040,034	0	21,040,034	23	-	5,303,267	0	5,303,267	23	-
N50 (bp)	3,177,341	0	3,906,945	23	-	670,370	0	748,892	23	-
N90 (bp)	247,663	0	483,079	23	-	70,261	0	129,142	23	-
NG50 (bp)	4,168,408	0	4,097,292	-	-	664,257	0	652,964	-	-
NG90 (bp)	965,081	0	799,088	-	-	62,351	0	29,319	-	-
Sample A1										
Total bp	1,249,218,340	0	1,143,554,647	851	105,663,693	1,115,757,570	0	1,014,043,497	115	101,714,073
Number	1,301	0	742	37	-	4,220	0	2,144	5	-
Max length (bp)	45,188,444	0	45,188,444	23	-	10,892,277	0	10,892,277	23	-
N50 (bp)	6,816,398	0	7,650,723	23	-	1,619,225	0	1,847,324	23	-
N90 (bp)	544,953	0	1,017,178	23	-	121,942	0	308,732	23	-
NG50 (bp)	8,633,890	0	8,633,890	-	-	1,713,272	0	1,713,272	-	-
NG90 (bp)	2,097,283	0	1,911,453	-	-	246,935	0	147,958	-	-
Sample A2										
Total bp	1,276,591,930	0	1,123,142,668	759	153,449,262	1,051,556,461	0	954,145,800	46	97,410,661
Number	3,111	0	1,861	33	-	8,500	0	5,550	2	-
Max length (bp)	12,728,721	0	12,728,721	23	-	1,850,512	0	1,850,512	23	-
N50 (bp)	1,547,009	0	1,860,711	23	-	259,207	0	291,680	23	-
N90 (bp)	155,020	0	265,534	23	-	49,549	0	80,915	23	-

NG50 (bp)	1,944,253	0	1,930,674	-	-	250,570	0	246,485	-	-
NG90 (bp)	451,290	0	348,627	-	-	39,495	0	-	-	-

d) Genomescope2.0 on HiFi raw data

Sample	Property	Homozygous (aa) (%)	Heterozygous (ab) (%)	Genome Haploid Length (bp)	Genome Repeat Length (bp)	Genome Unique Length (bp)	Model Fit (%)	Read Error Rate (%)
A1	min	98.8248	1.16067	1,060,090,691	148,917,848	911,172,844	85.263	0.184164
	max	98.8393	1.1752	1,060,747,433	149,010,104	911,737,328	98.2188	0.184164
	mean	98.83205	1.167935	1,060,419,062	148,963,976	911,455,086	91.7409	0.184164
A2	min	98.8456	1.13982	1,087,258,723	185,998,518	901,260,204	83.7977	0.234586
	max	98.8602	1.15442	1,088,738,793	186,251,716	902,487,077	99.4162	0.234586
	mean	98.8529	1.14712	1,087,998,758	186,125,117	901,873,641	91.60695	0.234586
2	min	98.7786	1.17519	1,080,524,880	165,703,559	914,821,322	84.3326	0.182921
	max	98.8248	1.22142	1,081,680,853	165,880,832	915,800,020	98.5606	0.182921
	mean	98.8017	1.198305	1,081,102,867	165,792,196	915,310,671	91.4466	0.182921
3	min	98.7622	1.21423	1,061,676,406	156,530,465	905,145,941	85.0648	0.170348
	max	98.7858	1.23777	1,062,843,484	156,702,535	906,140,949	98.8014	0.170348
	mean	98.774	1.226	1,062,259,945	156,616,500	905,643,445	91.9331	0.170348
4	min	98.8634	1.11572	1,078,140,346	172,221,680	905,918,667	84.7102	0.191115
	max	98.8843	1.13658	1,079,282,529	172,404,131	906,878,398	99.5474	0.191115
	mean	98.87385	1.12615	1,078,711,438	172,312,906	906,398,533	92.1288	0.191115

e) Purge_dups cutoffs

	2	3	4	A1	A2
kcov	12.3	9.81	10.1	16.5	7.47
value1	18.45	14.715	15.15	24.75	11.205
value2	55.35	44.145	45.45	74.25	33.615

f) Haplotig purging on HiFi assemblies

HAPLOTIG	170	115	129	67	243
HIGHCOV	0	0	0	0	1
JUNK	119	156	112	56	207
OVLP	84	92	112	76	142
REPEAT	619	710	742	472	831
TOT	992	1073	1095	671	1424
Removed (Mbp)	144	138	162	106	153

g) HiFi data alignment statistics

	A1	A2	2	3	4
Mapped reads	2,475,803	930,924	1,842,954	1,519,281	1,420,140
Alignments	2,994,011	1,215,608	2,258,506	1,825,799	1,777,646
Mapped bases	35,698,278,367	16,611,159,086	27,070,079,509	21,144,417,593	22,159,332,354
Mean mapped concordance	96.9233%	96.5095%	96.8043%	96.9713%	96.6645%
Max mapped read lenght	49,467	47,385	35,156	43,871	35,559
Mean mapped read lenght	11,923.20	13,664.90	11,985.80	11,580.90	12,465.50

Table S12 - Geographical origin, type of genomic data, population sizes and sequencing coverage (based on genome size) for all genomic data analysed. Related to Figure 3, S5 and S6. ds: dataset.

Identifier	Publication	Data type	N individuals	Location(s)	Subspecies	Average sequencing coverage
ds1	This publication	Pacbio HiFi WGS	5	Italy	*H. r. rustica*	~20x
ds2.1	Schield et al., 2021	Illumina WGS	34	China, Japan, Russia	*H. r. gutturalis*	~6.68x
ds2.2	Schield et al., 2021	Illumina WGS	25	China, Mongolia, Morocco, Russia	*H. r. rustica*	~6.08x
ds2.3	Schield et al., 2021	Illumina WGS	8	Israel	*H. r. transitiva*	~7.92x
ds2.4	Schield et al., 2021	Illumina WGS	10	Russia	*H. r. tytleri*	~6.35x
ds2.5	Schield et al., 2021	Illumina WGS	29	Mongolia, Russia (hybrid zone)	*H. r. gutturalis x H. r. tytleri*	~6.99x
ds2.6	Schield et al., 2021	Illumina WGS	21	China (hybrid zone)	*H. r. rustica x H. r. gutturalis*	~7.27x
ds2.7	Schield et al., 2021	Illumina WGS	16	Russia (hybrid zone)	*H. r. rustica x H. r. tytleri*	~5.79x
ds3.1.1	Smith et al., 2018	Illumina WGS	8	Egypt	*H. r. savignii*	~6.1x
ds3.1.2	Smith et al., 2018	Illumina WGS	8	Colorado, USA	*H. r. erythrogaster*	~7.83x
ds3.2.1	Smith et al., 2018	ddRAD	36	Egypt	*H. r. savignii*	~0.12x
ds3.2.2	Smith et al., 2018	ddRAD	26	Colorado, USA	*H. r. erythrogaster*	~0.08x
ds4	Scordato et al., 2017	ddRAD	533	Russia (west-to-east transect)	*H. r. rustica; H. r. tytleri; H. r. gutturalis*	~0.15x
ds5	Von Ronn et al., 2016	ddRAD	216	Transect from Sweden to Switzerland	*H. r. rustica*	~0.33x
ds6.1	Safran et al., 2016	ddRAD	145	Colorado, USA	*H. r. erythrogaster*	~0.02x
ds6.2	Safran et al., 2016	ddRAD	27	Ithaca, New York ,USA	*H. r. erythrogaster*	~0.03x
ds6.3	Safran et al., 2016	ddRAD	24	Czech Republic	*H. r. rustica*	~0.02x
ds6.4	Safran et al., 2016	ddRAD	16	Romania	*H. r. rustica*	~0.02x
ds6.5	Safran et al., 2016	ddRAD	50	Turkey	*H. r. rustica*	~0.02x
ds6.6	Safran et al., 2016	ddRAD	26	United Kingdom	*H. r. rustica*	~0.01x
ds6.7	Safran et al., 2016	ddRAD	45	Israel	*H. r. transitiva*	~0.02x
ds6.8	Safran et al., 2016	ddRAD	18	Taiwan	*H. r. gutturalis*	~0.03x

Table S16 - Pangenome ortholog analysis. **Related to Figure 5B and 5C.** The individual, or combination of individuals, are reported with the corresponding number of "private" or shared bHirRus1 genes. all = genes found in all individuals; bHirRus1 only = genes that were not found in the other individuals; HiFi raw-reads = 'bHirRus1 only' genes that were found in the HiFi raw reads; bHirRus1 only (no raw reads) = genes that were not found in the other individuals without the ones found in the HiFi raw reads.

Individuals	n. genes	Tot genes searched	%
all	16801	18136	92.64
bHirRus1 only	234	18136	1.29
HiFi raw reads	79	18136	0.44
bHirRus1 only (no raw reads)	155	18136	0.85
bHirRus1,2,3,4,A1	153	18136	0.84
bHirRus1,2,4,A1,A2	100	18136	0.55
bHirRus1,3,4,A1,A2	89	18136	0.49
bHirRus1,2,3,A1,A2	83	18136	0.46
bHirRus1,2,3,4,A2	67	18136	0.37
bHirRus1,2,4,A1	48	18136	0.26
bHirRus1,A1	47	18136	0.26
bHirRus1,2,3,A1	42	18136	0.23
bHirRus1,2,A1	41	18136	0.23
bHirRus1,2,A1,A2	41	18136	0.23
bHirRus1,3,4,A2	39	18136	0.22
bHirRus1,3	32	18136	0.18
bHirRus1,2	26	18136	0.14
bHirRus1,4	24	18136	0.13
bHirRus1,A2	23	18136	0.13
bHirRus1,A1,A2	23	18136	0.13
bHirRus1,2,3,4	19	18136	0.10
bHirRus1,3,4,A1	21	18136	0.12
bHirRus1,2,4	22	18136	0.12
bHirRus1,2,3,A2	21	18136	0.12
bHirRus1,2,4,A2	19	18136	0.10
bHirRus1,4,A1	18	18136	0.10
bHirRus1,3,A1,A2	16	18136	0.09
bHirRus1,3,A1	15	18136	0.08
bHirRus1,2,3	15	18136	0.08
bHirRus1,2,A2	12	18136	0.07
bHirRus1,4,A2	13	18136	0.07
bHirRus1,4,A1,A2	12	18136	0.07
bHirRus1,3,4	11	18136	0.06
bHirRus1,3,A2	8	18136	0.04

Table S18 - Dinucleotide content analysis in genes missing from both the HiFi assemblies and the raw data (HiFi missing genes) vs all bHirRus1 genes. Related to Figure 5E. a) Counts and percentages of sliding genomic windows (128 bp) exhibiting a specific dinucleotide content > 50% for the two groups of gene sets. GA and TC dinucleotides were added together to account for GA presence on both DNA strands. b), c), d) Chi-square test results performed on all three classes of dinucleotides. df: degrees of freedom.

a) >50%				
	HiFi missing genes		bHirRus1 genes	
	count	percentage	count	percentage
AT	100	0.54	67646	1.49
CG	483	2.59	40427	0.89
GA/TC	423	2.27	44452	0.98
tot windows number	18640		4532636	

b) Chi-square test (AT)			
	>50%	<50%	rows total
HiFi missing genes	100	18540	18640
bHirRus1 genes	67646	4464990	4532636
column total	67746	4483530	4551276

Expected values (AT)		
	>50%	<50%
HiFi missing genes	277.45745	18,362.54
bHirRus1 genes	67,468.543	4,465,167.5

Chi-square (AT)			
	>50%	<50%	Chi-square value
HiFi missing genes	113.499014	1.714966596	115.687786; df = 1; p < 0.0001
bHirRus1 genes	0.46675304	0.007052624	

c) Chi-square test (GC)			
	>50%	<50%	rows total
HiFi missing genes	483	18157	18640
bHirRus1 genes	40427	4492209	4532636
column total	40910	4510366	4551276

Expected values (GC)		
	>50%	<50%
HiFi missing genes	167.54914	18,472.451
bHirRus1 genes	40,742.451	4,491,893.5

Chi-square (GC)			
	>50%	<50%	Chi-square value
HiFi missing genes	593.910775	5.386899824	601.7622251; df = 1; p < 0.0001
bHirRus1 genes	2.44239706	0.022153072	

d) Chi-square test (GA/TC)			
	>50%	<50%	rows total
HiFi missing genes	423	18217	18640
bHirRus1 genes	44452	4488184	4532636
column total	44875	4506401	4551276

Expected values (GA/TC)		
	>50%	<50%
HiFi missing genes	183.78802	18,456.212
bHirRus1 genes	44,691.212	4,487,944.8

Chi-square (GA/TC)			
	>50%	<50%	Chi-square value
HiFi missing genes	311.349848	3.100439674	315.7434326; df =1; p < 0.0001
bHirRus1 genes	1.28039427	0.01275024	

Table S19 - Raw variants called with deepvariant from the alignment of the Hifi reads for the 5 barn swallow individuals against the linear reference genome in *camk2n2* coordinates. Related to Figure 5F and STAR methods. All 53 SNPs were found in the pangenome.

chr	pos	id	ref	alt	qual	filter	format	A1	A2	2	3	4
10	17272332	.	A	T	42.2	PASS	GT:GQ:DP:AD:VAF:PL	0/1:42:16:10,6:0.375:42,0,58	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17272523	.	G	A	53.6	PASS	GT:GQ:DP:AD:VAF:PL	0/1:37:16:11,5:0.3125:36,0,65	./.:.:.:.:.:.	0/1:54:18:9,9:0.5:53,0,71	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17272611	.	C	T	32.1	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:32:13:8,4:0.307692:32,0,57	./.:.:.:.:.:.
10	17272660	.	G	A	22.6	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:23:10:8,2:0.2:22,0,51	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17272666	.	C	T	46.2	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:31:10:2,8:0.8:46,0,31	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17273167	.	C	T	36.3	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:36:15:12,3:0.2:36,0,58
10	17273175	.	C	T	57.2	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:54:16:4,12:0.75:57,0,57
10	17273190	.	A	G	37	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:21:10:8,2:0.2:20,0,52	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:37:16:12,4:0.25:36,0,65
10	17273682	.	C	T	47.9	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:48:17:7,10:0.588235:47,0,59	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17273951	.	C	T	33.4	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:33:11:9,2:0.181818:33,0,54	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17273984	.	G	T	42.1	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:42:17:7,10:0.588235:42,0,55	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17273986	.	A	G	42.7	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:37:11:2,9:0.818182:42,0,38	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:42:16:5,11:0.6875:42,0,56
10	17274001	.	A	G	50.4	PASS	GT:GQ:DP:AD:VAF:PL	0/1:39:16:10,6:0.375:38,0,60	./.:.:.:.:.:.	./.:.:.:.:.:.	1/1:36:14:0,14:1:50,36,0	./.:.:.:.:.:.
10	17274004	.	T	C	44.3	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:44:16:4,12:0.75:44,0,60
10	17274096	.	C	T	40.1	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:40:17:10,7:0.411765:40,0,68	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274097	.	G	A	50.1	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:47:11:2,9:0.818182:50,0,50	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274102	.	G	A	45.9	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:43:11:2,9:0.818182:45,0,46	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274138	.	C	G	57.5	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:56:13:9,4:0.307692:57,0,61	./.:.:.:.:.:.
10	17274145	.	C	A	62.7	PASS	GT:GQ:DP:AD:VAF:PL	0/1:62:16:6,10:0.625:62,0,72	0/1:46:11:2,9:0.818182:46,0,53	./.:.:.:.:.:.	0/1:57:13:4,9:0.692308:58,0,62	./.:.:.:.:.:.
10	17274170	.	T	C	69.9	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:53:11:9,2:0.181818:53,0,58	1/1:69:17:0,17:1:69,78,0	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274178	.	G	T	54.2	PASS	GT:GQ:DP:AD:VAF:PL	0/1:54:16:10,6:0.375:54,0,70	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274184	.	C	T	55.6	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:55:13:4,9:0.692308:55,0,69	./.:.:.:.:.:.
10	17274218	.	G	A	52.6	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:51:11:2,9:0.818182:52,0,56	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274230	.	C	T	51.4	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	1/1:48:17:0,17:1:51,49,0	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274265	.	C	T	27.1	PASS	GT:GQ:DP:AD:VAF:PL	0/1:27:15:11,4:0.266667:27,0,50	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274355	.	G	A	69	PASS	GT:GQ:DP:AD:VAF:PL	1/1:63:16:0,16:1:68,63,0	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274536	.	T	G	56.4	PASS	GT:GQ:DP:AD:VAF:PL	0/1:50:14:9,5:0.357143:50,0,99	./.:.:.:.:.:.	1/1:44:17:0,16:0.941176:56,43,0	1/1:37:13:0,13:1:54,36,0	1/1:46:17:0,17:1:54,46,0
10	17274571	.	C	G	45.3	PASS	GT:GQ:DP:AD:VAF:PL	0/1:45:15:10,5:0.333333:45,0,70	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274605	.	G	A	51.9	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:39:11:2,9:0.818182:51,0,39	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274612	.	G	A	44	PASS	GT:GQ:DP:AD:VAF:PL	0/1:44:15:10,5:0.333333:44,0,70	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274615	.	G	A	55.2	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:49:11:2,9:0.818182:55,0,50	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274718	.	G	A	65.7	PASS	GT:GQ:DP:AD:VAF:PL	1/1:53:15:0,15:1:65,52,0	0/1:34:12:9,3:0.25:33,0,63	0/1:46:15:8,7:0.466667:46,0,74	0/1:60:13:4,9:0.692308:60,0,72	1/1:50:17:0,17:1:61,50,0
10	17274736	.	C	T	54.5	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:54:13:3,9:0.692308:54,0,76	0/1:44:16:5,11:0.6875:45,0,50
10	17274831	.	G	A	33.4	PASS	GT:GQ:DP:AD:VAF:PL	0/1:33:15:10,5:0.333333:33,0,59	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17274892	.	C	T	57.8	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:48:15:8,7:0.466667:47,0,72	0/1:58:13:4,9:0.692308:57,0,72	./.:.:.:.:.:.
10	17274985	.	G	C	49.7	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:49:12:3,9:0.75:49,0,59	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17275081	.	G	A	51.4	PASS	GT:GQ:DP:AD:VAF:PL	0/1:37:15:10,5:0.333333:36,0,56	./.:.:.:.:.:.	0/1:51:13:5,8:0.615385:51,0,62	./.:.:.:.:.:.	0/1:28:15:10,5:0.333333:27,0,55
10	17275090	.	G	A	46	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:46:14:7,7:0.5:46,0,63	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17275127	.	C	T	63.9	PASS	GT:GQ:DP:AD:VAF:PL	1/1:52:15:0,15:1:63,51,0	1/1:34:12:0,12:1:53,34,0	1/1:31:12:0,12:1:54,30,0	1/1:37:13:0,13:1:54,37,0	1/1:48:17:0,17:1:59,48,0
10	17275184	.	C	T	43.5	PASS	GT:GQ:DP:AD:VAF:PL	0/1:44:15:5,10:0.666667:43,0,64	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17275228	.	G	A	31.8	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:32:12:9,3:0.25:31,0,60	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17275259	.	T	C	28.8	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	0/1:29:12:9,3:0.25:28,0,60	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.

10	17275287	.	T	C	43.2	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:37:14:9,5:0.357143:36,0,66	0/1:43:17:12,5:0.294118:43,0,64
10	17275301	.	T	G	37.5	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:37:17:12,5:0.294118:37,0,57
10	17275417	.	C	A	33.7	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:34:14:9,5:0.357143:33,0,66	./.:.:.:.:.:.
10	17275426	.	C	A	49.9	PASS	GT:GQ:DP:AD:VAF:PL	0/1:35:15:5,10:0.666667:35,0,45	0/1:45:11:3,8:0.727273:46,0,50	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:48:17:5,12:0.705882:49,0,53
10	17275544	.	A	G	54.2	PASS	GT:GQ:DP:AD:VAF:PL	1/1:42:15:0,15:1:54,41,0	1/1:34:12:0,12:1:52,34,0	0/1:33:15:7,8:0.533333:33,0,53	1/1:40:14:0,14:1:53,40,0	1/1:46:18:0,18:1:50,48,0
10	17275565	.	T	G	57.1	PASS	GT:GQ:DP:AD:VAF:PL	1/1:38:14:0,14:1:57,38,0	1/1:30:11:0,11:1:52,29,0	1/1:51:14:0,14:1:56,51,0	1/1:38:14:0,14:1:56,38,0	1/1:42:16:0,16:1:53,41,0
10	17275775	.	G	A	57.9	PASS	GT:GQ:DP:AD:VAF:PL	1/1:46:16:0,16:1:57,46,0	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17275861	.	A	G	57.2	PASS	GT:GQ:DP:AD:VAF:PL	1/1:39:14:0,14:1:57,39,0	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.
10	17275916	.	C	T	53.4	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:53:14:9,5:0.357143:53,0,99	0/1:49:18:6,12:0.666667:50,0,55
10	17275933	.	C	A	51.1	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:51:18:6,12:0.666667:51,0,59
10	17275971	.	G	A	54.5	PASS	GT:GQ:DP:AD:VAF:PL	./.:.:.:.:.:.	./.:.:.:.:.:.	./.:.:.:.:.:.	0/1:54:14:9,5:0.357143:54,0,99	./.:.:.:.:.:.

Table S20 - Variants called from a chunk of the pangenome, representing *camk2n2* gene coordinates and the alignment of the 16 Illumina WGS barn swallow individuals. Related to Figure S7B and STAR methods. SNPs called as heterozygous with only one read supporting the alternate allele were not considered, for a more informative comparison with the variants set obtained with Freebayes using bHirRus1 as reference. The 20 SNPs deemed informative for the comparison are highlighted in yellow. The 8 SNPs found with Freebayes on the linear reference genomes are in red. The SNP that has no read support in the linear reference genome alignment is marked in blue (id: 125682575_8003823).

chrom	pos subgraph	pos real	id	ref	alt	qual	filter	info	format	sample
bHirRus1p.Chr10	142	17272294	8003816_125682588	A	T	21.7601	PASS	DP=4	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:4:3,1:-3.123771,-1.433656,-7.295518:16:-1.118820:2.721622:1
bHirRus1p.Chr10	147	17272299	125682588_125682590	C	A	19.3444	PASS	DP=4	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:4:3,1:-2.793988,-1.352093,-6.965735:14:-1.134127:4.820000:1
bHirRus1p.Chr10	178	17272330	8003821_125682575	T	A	17.2222	PASS	DP=5	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:5:4,1:-2.793988,-1.574314,-9.185783:15:-1.128317:4.820000:1
bHirRus1p.Chr10	180	17272332	125682575_8003823	A	C	36.9488	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:4,2:-4.713005,-1.495503,-9.185783:32:-1.099218:4.820000:2
bHirRus1p.Chr10	182	17272334	8003823_8003825	T	A	16.0578	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:5,1:-2.988551,-1.893444,-11.502741:10:-1.175890:4.820000:1
bHirRus1p.Chr10	192	17272344	125682388_125682390	G	T	16.8523	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:5,1:-3.212181,-2.031834,-11.726371:11:-1.162542:3.961020:1
bHirRus1p.Chr10	205	17272357	125682319_8003828	G	T	21.204	PASS	DP=4	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:4:3,1:-3.004886,-1.371591,-7.176633:16:-1.121613:3.102041:1
bHirRus1p.Chr10	231	17272383	125682365_125682367	T	A	21.204	PASS	DP=4	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:4:3,1:-3.004886,-1.371591,-7.176633:16:-1.121613:3.102041:1
bHirRus1p.Chr10	311	17272463	8003836_125682466	G	T	21.6333	PASS	DP=4	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:4:3,1:-3.094823,-1.417646,-7.266570:16:-1.119425:2.803921:1
bHirRus1p.Chr10	486	17272638	8003846_125682346	T	A	20.9534	PASS	DP=4	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:4:3,1:-2.958580,-1.350947,-7.130326:16:-1.122996:3.290859:1
bHirRus1p.Chr10	602	17272754	8003852_125682328	T	A	20.9534	PASS	DP=4	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:4:3,1:-2.958580,-1.350947,-7.130326:16:-1.122996:3.290859:1
bHirRus1p.Chr10	732	17272884	8003859_125682582	G	A	17.2222	PASS	DP=5	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:5:4,1:-3.089925,-1.870251,-9.256720:12:-1.157166:8.091813:1
bHirRus1p.Chr10	1554	17273706	8003872_125682311	T	C	8.61193	PASS	DP=9	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:9:8,1:-2.959867,-2.807141,-18.265868:1:-1.631308:6.715818:1
bHirRus1p.Chr10	1564	17273716	125682311_125682313	A	C	8.61193	PASS	DP=9	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:9:8,1:-2.851014,-2.698289,-18.096265:1:-1.631308:7.724138:1
bHirRus1p.Chr10	1577	17273729	125682313_125682315	C	T	73.8382	PASS	DP=9	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:9:5,4:-8.492711,-1.586000,-10.796985:69:-1.098612:7.326981:4
bHirRus1p.Chr10	1615	17273767	8003877_125682386	A	C	8.61193	PASS	DP=9	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:9:8,1:-2.930199,-2.777473,-18.222577:1:-1.631308:6.929824:1
bHirRus1p.Chr10	1815	17273967	8003885_125682557	A	C	10.6019	PASS	DP=8	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:8:7,1:-2.829511,-2.377900,-15.758486:4:-1.401306:8.080000:1
bHirRus1p.Chr10	1834	17273986	8003887_8003889	A	G	139.891	PASS	DP=8	GT:DP:AD:GL:GQ:GP:XD:MAD	1/0:8:1,7:-15.758486,-2.377900,-2.829511:4:-1.401306:8.080000:1
bHirRus1p.Chr10	1846	17273998	8003889_125682607	A	G	96.8595	PASS	DP=8	GT:DP:AD:GL:GQ:GP:XD:MAD	1/0:8:3,5:-10.741640,-1.532800,-6.399197:48:-1.098626:8.080000:3
bHirRus1p.Chr10	1892	17274044	8003893_125682580	A	T	16.301	PASS	DP=4	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:4:3,1:-3.477164,-2.355851,-7.648911:11:-1.171523:10.083898:1
bHirRus1p.Chr10	2016	17274168	8003905_125682592	A	C	15.5481	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:5,1:-2.885654,-1.845884,-11.399844:10:-1.185936:5.475000:1
bHirRus1p.Chr10	2018	17274170	125682592_8003907	T	C	81.6765	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	1/0:6:2,4:-9.138223,-1.447943,-4.665446:32:-1.099218:5.475000:2
bHirRus1p.Chr10	2203	17274355	8003919_8003921	G	A	368.98	PASS	DP=34	GT:DP:AD:GL:GQ:GP:XD:MAD	1/0:34:14,20:-38.992103,-2.571108,-25.274083:227:-1.098612:40.501129:14
bHirRus1p.Chr10	2384	17274536	125682678003931	T	G	576.229	PASS	DP=64	GT:DP:AD:GL:GQ:GP:XD:MAD	1/0:64:31,33:-59.606540,-2.460588,-55.048118:256:-1.098612:64.827583:31
bHirRus1p.Chr10	2463	17274615	8003937_8003939	G	A	8.70131	PASS	DP=56	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:56:49,7:-10.115984,-9.948140,-105.485570:1:-1.617078:52.156864:7
bHirRus1p.Chr10	2566	17274718	8003943_8003945	G	A	101.687	PASS	DP=7	GT:DP:AD:GL:GQ:GP:XD:MAD	1/0:7:2,5:-23.573326,-13.885219,-15.972590:20:-1.106757:48.821228:2
bHirRus1p.Chr10	2740	17274892	125682697125682629	C	T	37.0282	PASS	DP=30	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:30:25,5:-8.311884,-5.086438,-53.347652:45:-1.791789:29.160715:5
bHirRus1p.Chr10	2769	17274921	125682651_125682653	A	C	37.0282	PASS	DP=30	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:30:25,5:-8.311884,-5.086438,-53.347652:32:-1.099207:29.160715:5
bHirRus1p.Chr10	2771	17274923	125682653_125682655	A	C	26.2307	PASS	DP=26	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:26:22,4:-7.036796,-4.893954,-48.068006:28:-1.100051:29.160715:4
bHirRus1p.Chr10	2799	17274951	125682348_125682350	G	A	42.5883	PASS	DP=28	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:28:23,5:-8.407405,-4.625754,-48.897251:45:-1.098640:28.764706:5
bHirRus1p.Chr10	2807	17274959	125682352125682355	G	A	6.85465	PASS	DP=29	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:29:25,4:-5.338608,-5.549285,-54.716819:7:-1.943604:28.764706:3
bHirRus1p.Chr10	2825	17274977	125682355_125682357	T	A	14.9515	PASS	DP=30	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:30:26,4:-6.782842,-5.808605,-56.973967:9:-1.199464:28.764706:4
bHirRus1p.Chr10	2929	17275081	125682531_8003961	G	A	11.1889	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:5,1:-4.637202,-4.107901,-13.151392:5:-1.357584:17.735849:1
bHirRus1p.Chr10	2975	17275127	8003963_8003965	C	T	118.583	PASS	DP=5	GT:DP:AD:GL:GQ:GP:XD:MAD	1/1:5:0,5:-14.009115,-4.019978,-2.645934:13:-1.140006:14.268194:5
bHirRus1p.Chr10	3085	17275237	8003971_8003972	G	A	15.4585	PASS	DP=14	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:14:12,2:-4.335165,-3.305189,-26.978088:10:-1.187841:12.372264:2

bHirRus1p.Chr10	3107	17275259	8003972_8003974	T	C	15.4585	PASS	DP=14	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:14:12,2:-4.313470,-3.283494,-26.956393:10:-1.187841:12.820000:2
bHirRus1p.Chr10	3135	17275287	125682605_8003976	T	C	26.3073	PASS	DP=10	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:10:8,2:-4.433150,-2.282597,-17.964186:21:-1.105658:10.219090:2
bHirRus1p.Chr10	3149	17275301	8003976_8003978	T	G	12.5346	PASS	DP=7	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:7:6,1:-2.812155,-2.115361,-13.523680:6:-1.281771:7.618182:1
bHirRus1p.Chr10	3182	17275334	8003978_125682373	C	T	12.9371	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:5,1:-3.133077,-2.388370,-11.647267:7:-1.264134:10.800539:1
bHirRus1p.Chr10	3190	17275342	125682373_125682375	G	C	12.5346	PASS	DP=7	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:7:6,1:-2.800974,-2.104179,-13.550758:6:-1.281771:6.975758:1
bHirRus1p.Chr10	3194	17275346	125682375_125682377	T	G	12.5346	PASS	DP=7	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:7:6,1:-2.815686,-2.118891,-13.607429:6:-1.281771:6.333333:1
bHirRus1p.Chr10	3228	17275380	125682379_125682381	G	T	14.9722	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:5,1:-2.815686,-1.839164,-11.329875:9:-1.198960:6.333333:1
bHirRus1p.Chr10	3231	17275383	125682381_125682383	T	A	17.2222	PASS	DP=5	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:5:4,1:-2.858294,-1.638619,-9.131503:12:-1.157166:6.333333:1
bHirRus1p.Chr10	3253	17275405	125682558_125682560	G	C	12.8038	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:5,1:-3.196153,-2.467218,-11.710343:7:-1.269761:11.200000:1
bHirRus1p.Chr10	3259	17275411	125682562_125682564	G	C	12.8038	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:5,1:-3.196153,-2.467218,-11.710343:7:-1.269761:11.200000:1
bHirRus1p.Chr10	3261	17275413	125682564_125682566	G	A	12.8038	PASS	DP=6	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:6:5,1:-3.196153,-2.467218,-11.710343:7:-1.269761:11.200000:1
bHirRus1p.Chr10	3263	17275415	125682566_125682568	T	A	12.5346	PASS	DP=7	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:7:6,1:-3.196153,-2.499358,-13.740311:6:-1.281771:11.200000:1
bHirRus1p.Chr10	3392	17275544	125682578003990	A	G	622.135	PASS	DP=35	GT:DP:AD:GL:GQ:GP:XD:MAD	1/1:35:4,31:-68.424168,-7.269669,-6.819512:4:-1.402182:29.571428:31
bHirRus1p.Chr10	3413	17275565	125682426_8003992	T	G	823.566	PASS	DP=36	GT:DP:AD:GL:GQ:GP:XD:MAD	1/1:36:0,36:-84.018815,-12.097025,-2.139150:99:-1.098612:36.745098:36
bHirRus1p.Chr10	3764	17275916	125682464_8004005	C	T	34.868	PASS	DP=7	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:7:5,2:-4.682955,-1.673695,-11.361802:30:-1.099591:5.844156:2
bHirRus1p.Chr10	3860	17276012	8004012_125682452	T	A	17.5611	PASS	DP=4	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:4:3,1:-2.983987,-1.728462,-7.155734:12:-1.152653:7.403141:1
bHirRus1p.Chr10	4003	17276155	125682484_125682486	T	G	9.79399	PASS	DP=8	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:8:7,1:-2.855896,-2.517657,-15.784871:3:-1.476327:10.490196:1
bHirRus1p.Chr10	4015	17276167	125682486_125682488	T	A	12.5346	PASS	DP=7	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:7:6,1:-2.819654,-2.122860,-13.617009:6:-1.281771:6.252043:1
bHirRus1p.Chr10	4021	17276173	125682488_125682490	A	C	12.5346	PASS	DP=7	GT:DP:AD:GL:GQ:GP:XD:MAD	0/1:7:6,1:-4.422956,-3.726161,-15.712298:6:-1.281771:2.013889:1

Table S21 - Parameters used to filter vcf files of different publicly available datasets analyzed. Related to STAR methods. INFO/QUAL: quality score present in the INFO field of the vcf file. INFO/DP: sum of the read depth across all samples for a site. FMT/GQ: quality score in the format field for a genotype. FMT/DP: read depth at a particular site for a sample. Missingness fraction: fraction of missing genotypes at a site. MAF: minor allele frequency threshold. Discarded individuals: number of discarded individuals due to high number of missing genotypes.

identifier	info/qual	max info/dp	fmt/gq	max fmt/dp	min fmt/dp	missingness fraction	maf	discarded individuals
ds2.1	>30	1,367	>30	10	2	<0.1	>0.05	/
ds2.2	>30	1,369	>30	9	2	<0.1	>0.05	/
ds2.3	>30	1,364	>30	11	2	<0.1	>0.05	/
ds2.4	>30	1,368	>30	10	2	<0.1	>0.05	/
ds2.5	>30	1,368	>30	10	2	<0.1	>0.05	/
ds2.6	>30	1,367	>30	10	2	<0.1	>0.05	/
ds2.7	>30	1,368	>30	9	2	<0.1	>0.05	/
ds3.1.1	>30	180	>30	9	2	<0.1	>0.05	/
ds3.1.2	>30	179	>30	11	2	<0.1	>0.05	/
ds3.2.1	>30	703	>20	9	2	<0.2	>0.05	/
ds3.2.2	>30	727	>20	7	2	<0.2	>0.05	1
ds4	>30	8,598	>20	13	2	<0.2	>0.05	5
ds5	>30	12,852	>30	53	5	<0.3	>0.02	22
ds6.1	>20	770	>20	3	1	<0.4	>0.05	12
ds6.2	>20	770	>20	3	1	<0.4	>0.05	/
ds6.3	>20	770	>20	3	1	<0.4	>0.05	4
ds6.4	>20	770	>20	3	1	<0.4	>0.05	/
ds6.5	>20	770	>20	3	1	<0.4	>0.05	4
ds6.6	>20	770	>20	3	1	<0.4	>0.05	18
ds6.7	>20	770	>20	3	1	<0.4	>0.05	8
ds6.8	>20	770	>20	3	1	<0.4	>0.05	/

Table S22 - ROIs with highest average LD values (> 0.3) computed over 5 kbps windows using Illumina WGS data from ds3.1. Related to Figure S9, STAR methods and Data S1. A minimum of 100 markers per window was required to perform this analysis.

ROIn	chr	start	end	average r^2	marker pairs number	genes within the interval	Interval excluded for proximity to ambigous regions of the assembly	Interval excluded for not containing annotated genes
1	1	4,525,000	4,529,999	0.323718501	104	COL22A1		
2	1	26,525,000	26,529,999	0.326156141	190	/		X
3	1	96,335,000	96,339,999	0.485467547	105	CTDP1		
3	1	96,400,000	96,404,999	0.303938321	120	CTDP1		
3	1	96,450,000	96,454,999	0.447613366	120	CTDP1		
4	1	101,030,000	101,034,999	0.451958407	230	/		X
4	1	101,075,000	101,079,999	0.306657997	120	/		X
5	1	101,190,000	101,194,999	0.306817598	152	AMPH		
6	1	101,600,000	101,604,999	0.303188657	136	POU6F2		
7	1	103,225,000	103,229,999	0.426162424	105	MRPL32		
8	1	104,275,000	104,279,999	0.304933051	105	TPK1		
9	1	134,855,000	134,859,999	0.374239894	105	ABCB1		
9	1	134,860,000	134,864,999	0.327953806	120	ABCB1		
10	2	7,745,000	7,749,999	0.331257264	171	/		X
11	3	45,000	49,999	0.548404774	136	ACSS1		
11	3	80,000	84,999	0.331551673	231	TTBK1		
11	3	85,000	89,999	0.376619967	120	TTBK1		
11	3	90,000	94,999	0.410552098	120	TTBK1		
11	3	95,000	99,999	0.309422723	231	TTBK1		
11	3	105,000	109,999	0.417221146	120	TTBK1		
11	3	115,000	119,999	0.352963042	120	TTBK1		
11	3	160,000	164,999	0.329566118	105	SLC22A7		
11	3	195,000	199,999	0.391623229	190	POLR1B		
12	3	330,000	334,999	0.370706919	153	PPP2R5D		
12	3	355,000	359,999	0.363481778	253	MEA1, KLHDC3		
13	3	12,510,000	12,514,999	0.368875797	105	LOC120750516		
14	3	16,445,000	16,449,999	0.501605677	325	PCARE		
14	3	16,480,000	16,484,999	0.598724083	171	CLIP4		
14	3	16,485,000	16,489,999	0.462030338	171	CLIP4		
15	3	16,650,000	16,654,999	0.47011107	134	ALK		
15	3	16,715,000	16,719,999	0.538193851	276	ALK		
15	3	16,730,000	16,734,999	0.352739412	136	ALK		
16	3	16,850,000	16,854,999	0.553628128	171	LOC120750002		
16	3	16,880,000	16,884,999	0.387476236	105	/		X
16	3	16,890,000	16,894,999	0.460067064	171	/		X
16	3	16,895,000	16,899,999	0.330815482	105	/		X
16	3	16,905,000	16,909,999	0.310238401	210	/		X
16	3	16,910,000	16,914,999	0.350068784	105	/		X
16	3	16,920,000	16,924,999	0.598184708	105	YPEL5		
17	3	17,050,000	17,054,999	0.315353521	153	/		X
17	3	17,055,000	17,059,999	0.349880946	171	/		X
17	3	17,060,000	17,064,999	0.512528177	153	/		X
17	3	17,105,000	17,109,999	0.450686702	120	LCLAT1		
17	3	17,145,000	17,149,999	0.303347154	136	LCLAT1		
18	3	98,945,000	98,949,999	0.361068511	105	RPS7		
19	3	111,460,000	111,464,999	0.325996763	276	PLA2G7		
20	4	40,145,000	40,149,999	0.392746	103	PPP1R12A		
21	4	45,985,000	45,989,999	0.327447748	153	PLEKHG7		
22	5	21,565,000	21,569,999	0.434107134	105	/		X
23	5	36,835,000	36,839,999	0.313502134	105	LOC120753169		
24	5	68,095,000	68,099,999	0.302952662	136	/		X
25	6	27,755,000	27,759,999	0.334276961	120	MGA		
25	6	27,810,000	27,814,999	0.447479505	105	MAPKBP1		
25	6	27,845,000	27,849,999	0.452836549	120	MAPKBP1		
25	6	27,920,000	27,924,999	0.385319023	351	LOC120753633		
25	6	27,930,000	27,934,999	0.460366858	369	LOC120753633		
25	6	27,965,000	27,969,999	0.42517482	210	LOC120753633		
26	6	28,495,000	28,499,999	0.301443445	120	ZNF106		
27	6	28,825,000	28,829,999	0.353217755	153	UBR1		
28	6	29,215,000	29,219,999	0.332188077	105	RBM25		
29	6	29,825,000	29,829,999	0.312092979	210	/		X
29	6	29,835,000	29,839,999	0.365393453	595	/		X

29	6	29,940,000	29,944,999	0.503766325	104	SIPA1L1			
30	6	30,255,000	30,259,999	0.328386684	171	MAP3K9			
30	6	30,260,000	30,264,999	0.352607586	276	MAP3K9			
31	6	31,050,000	31,054,999	0.330805973	104	SLC39A9			
32	6	31,315,000	31,319,999	0.306403516	231	ACTN1			
33	6	32,305,000	32,309,999	0.328642463	300	GPHN			
34	6	32,430,000	32,434,999	0.300470827	119	GPHN			
34	6	32,455,000	32,459,999	0.362579067	105	GPHN			
35	6	36,325,000	36,329,999	0.329783717	119	/			X
36	6	36,960,000	36,964,999	0.41218069	105	/			X
37	6	37,280,000	37,284,999	0.469569797	253	NOVA1			
38	6	37,805,000	37,809,999	0.307179829	325	STXBP6			
38	6	37,810,000	37,814,999	0.578396897	153	STXBP6			
38	6	37,845,000	37,849,999	0.355639329	119	STXBP6			
38	6	37,855,000	37,859,999	0.354314546	171	STXBP6			
39	6	46,515,000	46,519,999	0.579355153	105	AP2A2			
39	6	46,535,000	46,539,999	0.503726853	135	AP2A2			
40	6	50,465,000	50,469,999	0.346252081	105	PLEKHA7			
41	6	50,760,000	50,764,999	0.339192987	120	SOX6			
42	6	51,280,000	51,284,999	0.386203126	171	/			X
43	6	52,020,000	52,024,999	0.323671792	406	LMO1			
44	6	52,165,000	52,169,999	0.312987042	153	LOC120754649			
44	6	52,260,000	52,264,999	0.314201757	105	LOC120754649			
44	6	52,270,000	52,274,999	0.372920521	120	LOC120754649			
44	6	52,275,000	52,279,999	0.372192095	135	LOC120754649			
44	6	52,315,000	52,319,999	0.337098639	105	/			X
44	6	52,400,000	52,404,999	0.315211782	741	TRIM66			
44	6	52,440,000	52,444,999	0.414994298	210	DENND2B			
44	6	52,455,000	52,459,999	0.312397507	120	DENND2B			
44	6	52,495,000	52,499,999	0.464522242	120	DENND2B			
45	6	53,680,000	53,684,999	0.771114866	136	/			X
45	6	53,695,000	53,699,999	0.587150949	153	/			X
45	6	53,705,000	53,709,999	0.705205103	105	/			X
45	6	53,725,000	53,729,999	0.872786758	120	CCDC34			
45	6	53,730,000	53,734,999	0.705027024	104	CCDC34			
45	6	53,735,000	53,739,999	0.740417802	105	CCDC34			
45	6	53,755,000	53,759,999	0.552005477	120	LGR4			
45	6	53,765,000	53,769,999	0.668553403	136	LGR4			
45	6	53,775,000	53,779,999	0.666817446	153	LGR4			
45	6	53,780,000	53,784,999	0.604254879	300	LGR4			
45	6	53,790,000	53,794,999	0.608621875	105	LGR4			
45	6	53,810,000	53,814,999	0.547712705	105	LGR4			
45	6	53,815,000	53,819,999	0.625098194	252	LGR4			
45	6	53,845,000	53,849,999	0.507142341	120	LIN7C			
45	6	53,860,000	53,864,999	0.30950362	190	/			X
45	6	53,870,000	53,874,999	0.349581369	153	/			X
45	6	53,875,000	53,879,999	0.377250405	378	/			X
45	6	53,885,000	53,889,999	0.432021741	190	BDNF			
45	6	53,895,000	53,899,999	0.496976461	120	BDNF			
45	6	53,900,000	53,904,999	0.329262907	120	BDNF			
45	6	53,910,000	53,914,999	0.34897478	120	BDNF			
45	6	53,945,000	53,949,999	0.37879887	324	/			X
45	6	53,950,000	53,954,999	0.419861929	153	/			X
46	8	525,000	529,999	0.756105245	120	LOC120755920			
47	8	2,910,000	2,914,999	0.323768569	136	MCU			
48	8	3,210,000	3,214,999	0.327505282	136	LOC120756098			
49	10	1,300,000	1,304,999	0.312055501	153	VAMP2			
50	11	16,830,000	16,834,999	0.430271256	276	/	X		
50	11	16,845,000	16,849,999	0.460269469	210	/			X
50	11	16,855,000	16,859,999	0.428974743	231	/			X
51	11	18,220,000	18,224,999	0.546459528	171	/			X
51	11	18,235,000	18,239,999	0.551595553	120	/			X
51	11	18,260,000	18,264,999	0.773465513	136	/			X
51	11	18,270,000	18,274,999	0.9384316	190	/			X
51	11	18,275,000	18,279,999	0.694954171	105	/			X
51	11	18,295,000	18,299,999	0.318043332	153	/			X
51	11	18,365,000	18,369,999	0.525914905	120	/			X
51	11	18,410,000	18,414,999	0.336137998	561	/			X
51	11	18,515,000	18,519,999	0.463150477	105	/			X
51	11	18,595,000	18,599,999	0.32001957	120	/			X
52	11	19,455,000	19,459,999	0.331963055	595	BEAN1			

52	11	19,500,000	19,504,999	0.300486863	435	LOC120757948			
52	11	19,510,000	19,514,999	0.333400554	190	TK2			
52	11	19,535,000	19,539,999	0.382415971	136	LOC120757604,CMTM3	X		
52	11	19,550,000	19,554,999	0.393890105	171	/			X
52	11	19,560,000	19,564,999	0.390796904	300	CMTM4			
52	11	19,565,000	19,569,999	0.381078004	190	CMTM4			
52	11	19,575,000	19,579,999	0.346412034	300	CMTM4			
52	11	19,580,000	19,584,999	0.495238444	351	CMTM4			
52	11	19,585,000	19,589,999	0.509732531	151	CMTM4			
52	11	19,590,000	19,594,999	0.463295287	187	/			X
52	11	19,595,000	19,599,999	0.504887215	120	DYNC1LI2			
52	11	19,605,000	19,609,999	0.713151153	171	DYNC1LI2			
52	11	19,610,000	19,614,999	0.705062508	190	DYNC1LI2			
52	11	19,615,000	19,619,999	0.560123309	190	DYNC1LI2			
53	11	19,800,000	19,804,999	0.947497399	153	LOC120757950	X		
53	11	19,815,000	19,819,999	0.696842914	168	LOC120757727			
53	11	19,915,000	19,919,999	0.831808	253	LOC120757950			
53	11	19,925,000	19,929,999	0.753129214	231	LRP3			
53	11	19,930,000	19,934,999	0.611347163	190	LRP3			
53	11	19,940,000	19,944,999	0.403813532	595	LRP3			
53	11	19,945,000	19,949,999	0.596263267	253	LRP3			
53	11	19,965,000	19,969,999	0.696903778	171	SLC7A10			
53	11	19,970,000	19,974,999	0.531826285	136	SLC7A10			
53	11	19,980,000	19,984,999	0.317947238	351	SLC7A10			
53	11	19,990,000	19,994,999	0.336530293	231	SLC7A10			
53	11	19,995,000	19,999,999	0.308402149	698	SLC7A10			
53	11	20,000,000	20,004,999	0.477269596	561	SLC7A10			
53	11	20,005,000	20,009,999	0.308389392	136	SLC7A10			
53	11	20,020,000	20,024,999	0.385645733	210	/			X
53	11	20,025,000	20,029,999	0.321230946	231	/			X
53	11	20,035,000	20,039,999	0.413027229	171	/			X
53	11	20,040,000	20,044,999	0.376046977	190	/			X
53	11	20,045,000	20,049,999	0.569983013	136	/			X
53	11	20,050,000	20,054,999	0.443500276	190	/			X
53	11	20,055,000	20,059,999	0.387810415	136	/			X
54	11	21,200,000	21,204,999	0.365741393	435	UBA2			
55	11	21,385,000	21,389,999	0.496061776	153	PDP2			
55	11	21,390,000	21,394,999	0.312506656	231	CA7			
55	11	21,400,000	21,404,999	0.59924966	120	CA7	X		
55	11	21,405,000	21,409,999	0.827014714	630	NAE1	X		
56	12	9,430,000	9,434,999	0.312684787	405	EEFSEC			
57	16	14,720,000	14,724,999	0.344505056	105	NDRG3			
57	16	14,740,000	14,744,999	0.37214136	253	NDRG3			
57	16	14,755,000	14,759,999	0.320529563	136	NDRG3			
57	16	14,770,000	14,774,999	0.354718301	136	NDRG3			
58	17	160,000	164,999	0.369956025	105	/			X
59	18	6,860,000	6,864,999	0.340840058	136	/			X
60	27	20,000	24,999	0.429871942	105	/			X
61	W	5,120,000	5,124,999	0.689897033	120	MBNL3			
61	W	5,125,000	5,129,999	0.643972838	120	MBNL3			
61	W	5,130,000	5,134,999	0.612508051	136	MBNL3			
62	W	9,325,000	9,329,999	0.46989773	105	/			X
63	W	9,570,000	9,574,999	0.697001733	105	RBMX			
63	W	9,605,000	9,609,999	0.666389717	120	LOC120764854			
64	W	12,075,000	12,079,999	0.528482703	171	ZC3H12B			
65	W	12,290,000	12,294,999	0.558123099	171	LOC120764947			
66	W	19,310,000	19,314,999	0.440821923	136	POF1B			
67	W	19,460,000	19,464,999	0.60091764	136	LOC120764804			
68	W	20,630,000	20,634,999	0.467206752	136	/			X
69	W	23,465,000	23,469,999	0.462534376	276	ZBTB33			
69	W	23,520,000	23,524,999	0.580058622	231	UPF3B,RPL39,LOC120765016			
69	W	23,615,000	23,619,999	0.533652241	153	SEPTIN6			
69	W	23,655,000	23,659,999	0.484432125	496	NKRF			
70	W	28,315,000	28,319,999	0.837218083	120	ATP11C			
71	W	31,595,000	31,599,999	0.579872077	210	LOC120764863			
72	Z	58,080,000	58,084,999	0.433165801	120	/			X
72	Z	58,115,000	58,119,999	0.322064155	351	/			X
72	Z	58,120,000	58,124,999	0.453374826	276	/			X
72	Z	58,125,000	58,129,999	0.399757528	351	/			X
72	Z	58,145,000	58,149,999	0.304713724	465	/			X
72	Z	58,175,000	58,179,999	0.419082732	153	/			X

73	Z	59,570,000	59,574,999	0.368467485	276	/	X
73	Z	59,575,000	59,579,999	0.312210827	120	YTHDC2	
74	Z	77,225,000	77,229,999	0.318757035	171	/	X
75	Z	82,775,000	82,779,999	0.304575576	325	CUL4B	
76	Z	86,270,000	86,274,999	0.433500256	120	LOC120765675	
77	Z	86,550,000	86,554,999	0.370511446	104	DACH2	
78	Z	89,970,000	89,974,999	0.569778107	104	/	X
78	Z	89,975,000	89,979,999	0.63959112	300	/	X
78	Z	90,010,000	90,014,999	0.698744137	153	/	X

Table S23 - Missing Chelidonia BUSCOs that were found in bHirRus1. Related to Data S1. Information about BUSCO genes in bHirRus1 were retrieved from the "full_table.tsv" BUSCO output.

Busco ID	status in Chelidonia	status in bHirRus1	bHirRus1 chr.	Gene Start	Gene End	Score	Lenght	OrthoDB url	Description
129214at7742	Missing	Complete	SUPER_2	43763683	43823851	1055.5	540	https://www.orthodb.org/v10?query=129214at7742	probable ATP-dependent RNA helicase DDX10
239747at7742	Missing	Complete	SUPER_29_unloc_1	53146	55386	462.2	352	https://www.orthodb.org/v10?query=239747at7742	Ribosomal RNA adenine dimethylase domain containing1
290517at7742	Missing	Complete	scaffold_270_arrow_ctg1	31546	39663	364.8	236	https://www.orthodb.org/v10?query=290517at7742	eukaryotic translation initiation factor 3 subunit G
295275at7742	Missing	Complete	SUPER_Z	38438718	38440601	465.2	326	https://www.orthodb.org/v10?query=295275at7742	recQ-mediated genome instability protein 1
323438at7742	Missing	Complete	SUPER_Z	53288635	53294590	419	235	https://www.orthodb.org/v10?query=323438at7742	metallo-beta-lactamase domain-containing protein 2
368989at7742	Missing	Complete	SUPER_Z	15924176	15931677	351	192	https://www.orthodb.org/v10?query=368989at7742	Transmembrane protein 267
424874at7742	Missing	Complete	SUPER_Z	64307048	64309376	187.5	89	https://www.orthodb.org/v10?query=424874at7742	Ubiquitin-like protein ATG12
46657at7742	Missing	Complete	SUPER_11	15250903	15268907	1840.8	917	https://www.orthodb.org/v10?query=46657at7742	teashirt homolog 3
73217at7742	Missing	Complete	SUPER_15	1638202	1645187	1250.1	626	https://www.orthodb.org/v10?query=73217at7742	WD repeat domain 24
77322at7742	Missing	Complete	SUPER_8	16206525	16227942	1266.6	690	https://www.orthodb.org/v10?query=77322at7742	lymphoid-specific helicase
80213at7742	Missing	Complete	SUPER_Z	46971972	47003188	1213.8	758	https://www.orthodb.org/v10?query=80213at7742	WD repeat-containing protein 36
139811at7742	Missing	Fragmented	SUPER_17	7998088	8025201	373.8	283	https://www.orthodb.org/v10?query=139811at7742	Mitogen-activated protein kinase
36859at7742	Missing	Fragmented	scaffold_622_arrow_ctg1	908	8003	558.3	276	https://www.orthodb.org/v10?query=36859at7742	symplekin

Table S30 – Mitogenome annotation. Related to Figure S1 and Data S1.

Annotation	start	end	e-value	strand
OH	23	229	1.7E+06	+
trnF(gaa)	248	317	1.1E-11	+
rrnS	316	1290	0	+
trnV(tac)	1289	1359	9.4E-12	+
rrnL	1367	2943	0	+
trnL2(taa)	2967	3042	8.7E-13	+
nad1	3069	4040	8.9E+08	+
trnI(gat)	4054	4125	1.7E-08	+
trnQ(ttg)	4132	4204	1.9E-14	-
trnM(cat)	4203	4272	2.1E-13	+
nad2	4272	5313	5.9E+08	+
trnW(tca)	5312	5383	1.8E-14	+
trnA(tgc)	5384	5453	5.4E-12	-
trnN(gtt)	5462	5535	4.8E-13	-
trnC(gca)	5535	5601	4.9E-10	-
trnY(gta)	5600	5671	1.5E-14	-
cox1	5672	7229	2.4E+09	+
trnS2(tga)	7220	7295	4.5E-14	-
trnD(gtc)	7299	7368	4.5E-12	+
cox2	7378	8062	5.0E+08	+
trnK(ttt)	8063	8134	2.5E-13	+
atp8	8135	8303	5.3E+06	+
atp6	8293	8977	3.2E+08	+
cox3	8983	9768	7.3E+08	+
trnG(tcc)	9767	9836	5.3E-12	+
nad3	9836	10187	9.2E+07	+
trnR(tcg)	10189	10259	8.6E-12	+
nad4l	10260	10557	4.8E+07	+
nad4	10550	11928	1.4E+09	+
trnH(gtg)	11928	11997	1.9E-11	+
trnS1(gct)	11997	12064	2.9E-09	+
trnL1(tag)	12063	12134	3.4E-19	+
nad5	12134	13952	2.0E+09	+
cob	13960	15103	1.6E+09	+
trnT(tgt)	15107	15176	3.5E-12	+
OL	15862	15896	3.6E-02	-

Table S31 - 38 species with *camk2n2* transcripts downloaded from NCBI. Related to STAR methods and Data S1. The last column regards the level of association with humans. W = wild; D = domestic; S = synanthropic.

Gene symbol	Gene ID	Description	Scientific name	Common name	RefSeq Transcript accessions	RefSeq Protein accessions	Human association
camk2n2	101747380	calcium/calmodulin dependent protein kinase II inhibitor 2	*Gallus gallus*	Chicken	XM_015291551.3	XP_015147037.1	D
camk2n2	102107607	calcium/calmodulin dependent protein kinase II inhibitor 2	*Pseudopodoces humilis*	Tibetan ground-tit	XM_005525032.1	XP_005525089.1	W
camk2n2	103528144	calcium/calmodulin dependent protein kinase II inhibitor 2	*Calypte anna*	Anna's hummingbird	XM_030456215.1	XP_030312075.1	W
camk2n2	104057870	calcium/calmodulin dependent protein kinase II inhibitor 2	*Cuculus canorus*	Common cuckoo	XM_009559309.1	XP_009557604.1	W
camk2n2	105759523	calcium/calmodulin dependent protein kinase II inhibitor 2	*Taeniopygia guttata*	Zebra finch	XM_041718084.1	XP_041574018.1	D
camk2n2	106850415	calcium/calmodulin dependent protein kinase II inhibitor 2	*Sturnus vulgaris*	Common starling	XM_014871707.1	XP_014727193.1	S
camk2n2	107208613	calcium/calmodulin dependent protein kinase II inhibitor 2	*Parus major*	Great Tit	XM_015637111.2	XP_015492597.1	S
camk2n2	107318230	calcium/calmodulin dependent protein kinase II inhibitor 2	*Coturnix japonica*	Japanese quail	XM_015871828.2	XP_015727314.1	D
camk2n2	108495276	calcium/calmodulin dependent protein kinase II inhibitor 2	*Lepidothrix coronata*	Blue-crowned manakin	XM_017810817.1	XP_017666306.1	W
camk2n2	110397805	calcium/calmodulin dependent protein kinase II inhibitor 2	*Numida meleagris*	Helmeted guineafowl	XM_021394645.1	XP_021250320.1	D
camk2n2	110477863	calcium/calmodulin dependent protein kinase II inhibitor 2	*Lonchura striata domestica*	Bengalese finch	XM_021544020.2	XP_021399695.1	D
camk2n2	113844458	calcium/calmodulin dependent protein kinase II inhibitor 2	*Anas platyrhynchos*	Mallard	XM_027464230.2	XP_027320031.1	D
camk2n2	113949204	calcium/calmodulin dependent protein kinase II inhibitor 2	*Corapipo altera*	White-ruffed manakin	XM_027647498.1	XP_027503299.1	W
camk2n2	113968780	calcium/calmodulin dependent protein kinase II inhibitor 2	*Neopelma chrysocephalum*	Saffron-crested tyrant-manakin	XM_027685656.1	XP_027541457.1	W
camk2n2	114003775	calcium/calmodulin dependent protein kinase II inhibitor 2	*Pipra filicauda*	Wire-tailed manakin	XM_027751109.2	XP_027606910.1	W
camk2n2	114060775	calcium/calmodulin dependent protein kinase II inhibitor 2	*Empidonax traillii*	Willow flycatcher	XM_027890551.1	XP_027746352.1	W
camk2n2	115347171	calcium/calmodulin dependent protein kinase II inhibitor 2	*Aquila chrysaetos chrysaetos*	Goled eagle	XM_030028241.2	XP_029884101.1	W
camk2n2	115611213	calcium/calmodulin dependent protein kinase II inhibitor 2	*Strigops habroptila*	Kakapo	XM_030493772.1	XP_030349632.1	W
camk2n2	115907087	calcium/calmodulin dependent protein kinase II inhibitor 2	*Camarhynchus parvulus*	Small tree finch	XM_030954740.1	XP_030810600.1	W
camk2n2	116235402	calcium/calmodulin dependent protein kinase II inhibitor 2	*Phasianus colchicus*	Ring-necked pheasant	XM_031603550.1	XP_031459410.1	D
camk2n2	116448661	calcium/calmodulin dependent protein kinase II inhibitor 2	*Corvus moneduloides*	New Caledonian crow	XM_032119428.1	XP_031975319.1	W

camk2n2	116492425	calcium/calmodulin dependent protein kinase II inhibitor 2	*Aythya fuligula*	Tufted duck	XM_032193158.1	XP_032049049.1	W
camk2n2	116791811	calcium/calmodulin dependent protein kinase II inhibitor 2	*Chiroxiphia lanceolata*	Lance-tailed manakin	XM_032698173.1	XP_032554064.1	W
camk2n2	117000567	calcium/calmodulin dependent protein kinase II inhibitor 2	*Catharus ustulatus*	Swainson's thrush	XM_033068263.1	XP_032924154.1	W
loc117436255	117436255	calcium/calmodulin dependent protein kinase II inhibitor 2	*Melopsittacus undulatus*	Budgerigar	XM_034063749.1	XP_033919640.1	D
camk2n2	118157648	calcium/calmodulin dependent protein kinase II inhibitor 2	*Oxyura jamaicensis*	Ruddy duck	XM_035312059.1	XP_035167950.1	W
camk2n2	118247211	calcium/calmodulin dependent protein kinase II inhibitor 2	*Cygnus atratus*	Black swan	XM_035545010.1	XP_035400903.1	S
camk2n2	118690048	calcium/calmodulin dependent protein kinase II inhibitor 2	*Molothrus ater*	Brown-headed cowbird	XM_036388723.1	XP_036244616.1	W
camk2n2	119157061	calcium/calmodulin dependent protein kinase II inhibitor 2	*Falco rusticolus*	Gyrfalcon	XM_037407595.1	XP_037263492.1	W
camk2n2	119704357	calcium/calmodulin dependent protein kinase II inhibitor 2	*Motacilla alba alba*	White wagtail	XM_038145455.1	XP_038001383.1	S
camk2n2	120410472	calcium/calmodulin dependent protein kinase II inhibitor 2	*Corvus cornix cornix*	Hooded crow	XM_039557048.1	XP_039412982.1	S
camk2n2	120504284	calcium/calmodulin dependent protein kinase II inhibitor 2	*Passer montanus*	Eurasian tree sparrow	XM_039713643.1	XP_039569577.1	S
camk2n2	120757412	calcium/calmodulin dependent protein kinase II inhibitor 2	*Hirundo rustica*	Barn swallow	XM_040074724.1	XP_039930658.1	S
camk2n2	121075074	calcium/calmodulin dependent protein kinase II inhibitor 2	*Cygnus olor*	Mute swan	XM_040567937.1	XP_040423871.1	S
camk2n2	121096835	calcium/calmodulin dependent protein kinase II inhibitor 2	*Falco naumanni*	Lesser kestrel	XM_040613367.1	XP_040469301.1	S
camk2n2	121336579	calcium/calmodulin dependent protein kinase II inhibitor 2	*Onychostruthus taczanowskii*	White-rumped snowfinch	XM_041405897.1	XP_041261831.1	W
camk2n2	121353747	calcium/calmodulin dependent protein kinase II inhibitor 2	*Pyrgilauda ruficollis*	Rufous-necked snowfinch	XM_041467857.1	XP_041323791.1	W
camk2n2	121671265	calcium/calmodulin dependent protein kinase II inhibitor 2	*Corvus kubaryi*	Mariana crow	XM_042042637.1	XP_041898571.1	W

Table S33 - Number of SNPs identified in the different datasets and summary of the values relative to variant depth in the different datasets after our filtering pipeline. Related to Data S1. For each dataset, excluding the column 'SNP number after filtering', the first value refers to average depth of coverage per site, while the second refers to the mean depth of coverage per individual of the dataset.

Identifier	SNP number after filtering	min	1st quantile	median	mean	3rd quantile	max
ds2.1	4,180,839	2.3 ; 3.09	4.58 ; 4.84	5.21 ; 5.16	5.21 ; 5.21	5.82 ; 5.71	11.85 ; 6.54
ds2.2	2,296,850	2.09 ; 2.04	4.25 ; 4.43	4.76 ; 4.86	4.77 ; 4.75	5.28 ; 5.25	7.84 ; 6.14
ds2.3	5,783,842	1.87 ; 5.25	4.87 ; 5.41	5.75 ; 5.66	5.71 ; 5.71	6.5 ; 5.88	10.5 ; 6.55
ds2.4	7,543,250	1.8 ; 2.98	4.3 ; 4.95	5.1 ; 5.06	5.14 ; 5.13	5.9 ; 5.29	10.2 ; 6.66
ds2.5	3,762,802	2.29 ; 3.27	4.64 ; 4.89	5.24 ; 5.28	5.25 ; 5.25	5.86 ; 5.73	8.65 ; 7.11
ds2.6	6,471,459	2 ; 4.7	4.57 ; 4.93	5.38 ; 5.46	5.37 ; 5.37	6.14; 5.57	10.57 ; 6.17
ds2.7	2,414,350	2 ; 2.43	4.12 ; 4.51	4.67 ; 4.63	4.68 ; 4.67	5.25 ; 4.74	9 ; 6.97
ds3.1.1	3,111,728	2 ; 4.08	4.25 ; 4.75	4.87 ; 4.91	4.87 ; 4.87	5.5 ; 5.08	8.62 ; 5.28
ds3.1.2	5,022,964	2 ; 4.91	5 ; 5.49	5.87 ; 5.88	5.8 ; 5.8	6.62 ; 6.24	10.62 ; 6.29
ds3.2.1	42,022	2.35 ; 2.78	3.77 ; 3.62	4.36 ; 4.35	4.5 ; 4.49	5.11 ; 5.27	7.97 ; 6.72
ds3.2.2	16,103	2.21 ; 2.41	3.3 ; 3.24	3.76 ; 4.15	3.88 ; 3.87	4.36 ; 4.45	6.68 ; 4.96
ds4	30,491	3.63 ; 1.76	4.94 ; 3.89	5.63 ; 5.32	5.8 ; 5.73	6.53 ; 7.13	12.78 ; 14.14
ds5	28,840	7.17 ; 7.96	19.71 ; 21.09	26.19 ; 23.95	27.25 ; 27.25	33.93 ; 28.33	64.41 ; 101.38
ds6.1	22,650	1.34 ; 1.25	1.71 ; 1.53	1.85 ; 1.72	1.89 ; 1.83	2.02 ; 2.09	3.09 ; 3.06
ds6.2	36,835	1.06 ; 1.45	1.71 ; 1.7	1.92 ; 1.87	1.96 ; 1.95	2.17 ; 2.12	3.81 ; 2.74
ds6.3	26,291	1 ; 1.24	1.65 ; 1.49	1.86 ; 1.63	1.9 ; 1.83	2.08 ; 2.07	7.59 ; 3.03
ds6.4	17,925	1 ; 1.27	1.4 ; 1.48	1.6 ; 1.59	1.66 ; 1.64	1.83 ; 1.81	3.21 ; 2.12
ds6.5	15,682	1.14 ; 1.23	1.59 ; 1.47	1.76 ; 1.66	1.79 ; 1.74	1.94 ; 1.88	3.13 ; 3.32
ds6.6	18,446	1 ; 1.32	1.4 ; 1.43	1.67 ; 1.57	1.71 ; 1.67	2 ; 1.64	3.86 ; 2.66
ds6.7	16,824	1.04 ; 1.24	1.56 ; 1.53	1.71 ; 1.72	1.74 ; 1.71	1.89 ; 1.87	4.23 ; 2.59
ds6.8	31,804	1 ; 1.4	1.61 ; 1.62	1.85 ; 1.78	1.88 ; 1.86	2.08 ; 2.07	4.64 ; 2.74

Table S35 - Correlation between average LD values and distance from chromosome ends in the different datasets. Related to Figure S11 and Data S1. The correlation between LD values (estimated as r2) and distance from chromosome ends was computed with the Spearman nonparametric rank test. Marker pairs were grouped using 10kb as distance bin value from chromosome ends.

	macrochromosomes	**intermediate chromosomes**	**microchromosomes**
***H. r. savignii* (ds3.1.1)**	rho = 0.14, S =6.79e+10, p-value < 2.2e-16	rho = 0.17, S =983239900, p-value = 5.385e-14	rho = 0.03, S =134304856, p-value = 0.3134
***H. r. erythrogaster* (ds3.1.2)**	rho = 0.17, S=6.57e+10, p-value < 2.2e-16	rho = 0.36, S =754182052, p-value < 2.2e-16	rho = -0.02 , S =141576406, p-value = 0.5508
***H. r. gutturalis* (ds2.1)**	rho = 0.19, S =6.39e+10, p-value < 2.2e-16	rho = 0.25, S =888761490, p-value < 2.2e-16	rho = 0.27, S =100906066, p-value < 2.2e-16
***H. r. rustica* (ds2.2)**	rho = -0.27, S =1e+11, p-value < 2.2e-16	rho = 0.2, S =948966740, p-value < 2.2e-16	rho = -0.07 , S =148490576, p-value = 0.03367

References

1. Formenti, G., Chiara, M., Poveda, L., Francoijs, K.-J., Bonisoli-Alquati, A., Canova, L., Gianfranceschi, L., Horner, D.S., and Saino, N. (2019). SMRT long reads and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (Hirundo rustica rustica). Gigascience 8. 10.1093/gigascience/giy142.

2. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V, and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212. 10.1093/bioinformatics/btv351.

3. Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V, and Zdobnov, E.M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35, 543–548. 10.1093/molbev/msx319.

4. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In Methods in Molecular Biology, pp. 227–2

5. Feron, R., Zahm, M., Cabau, C., Klopp, C., Roques, C., Bouchez, O., Eché, C., Valière, S., Donnadieu, C., Haffray, P., et al. (2020). Characterization of a Y-specific duplication/insertion of the anti-Mullerian hormone type II receptor gene based on a chromosome-scale genome assembly of yellow perch, Perca flavescens. Mol. Ecol. Resour. 20, 531–543. 10.1111/1755-0998.13133.

6. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36, 2896–2898. 10.1093/bioinformatics/btaa025.

7. Formenti, G., Rhie, A., Balacco, J., Haase, B., Mountcastle, J., Fedrigo, O., Brown, S., Capodiferro, M.R., Al-Ajli, F.O., Ambrosini, R., et al. (2021). Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome Biol. 22, 120. 10.1186/s13059-021-02336-9.

8. Lombardo, G., Rambaldi Migliore, N., Colombo, G., Capodiferro, M.R., Formenti, G., Caprioli, M., Moroni, E., Caporali, L., Lancioni, H., Secomandi, S., et al. (2022). The Mitogenome Relationships and Phylogeography of Barn Swallows (Hirundo rustica). Mol. Biol. Evol. 39. 10.1093/molbev/msac113.

9. Bulatova, N. SH., EN PANOV, AND SI RADZHABLI. 1971. Description of karyotypes of some species of birds of the USSR fauna. Proc. Acad. Sci. USSR 199, 1420-1423 LB-M2N0x.

10. Malinovskaya, L.P., Tishakova, K., Shnaider, E.P., Borodin, P.M., and Torgasheva, A.A. (2020). Heterochiasmy and Sexual Dimorphism: The Case of the Barn Swallow (Hirundo rustica, Hirundinidae, Aves). Genes (Basel). 11. 10.3390/genes11101119.

11. Barcellos, S.A., Kretschmer, R., Souza, M.S. de, Costa, A.L., Degrandi, T.M., Lopes, C.F., Ferguson-Smith, M.A., Pereira, J., Oliveira, E.H.C. de, Gunski, R.J., et al. (2020). Comparative analyses of three swallow species (Aves, Passeriformes, Hirundinidae): Insights on karyotype evolution and genomic organization. Genet. Mol. Biol. 43, e20190232. 10.1590/1678-4685-gmb-2019-0232.

12. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. 11, 1432. 10.1038/s41467-020-14998-3.

13. Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21, 245. 10.1186/s13059-020-02134-9.

14. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. Genome Res. 27, 722–736. 10.1101/gr.215087.116.

15. Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods 13, 1050–1054. 10.1038/nmeth.4035.

16. Morgulis, A., Gertz, E.M., Schaffer, A.A., and Agarwala, R. (2006). WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22, 13

17. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Curr. Protoc. Bioinforma. 25, Unit 4.10. 10.1002/0471250953.bi0410s25.

18. Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A.H., Xie, D., Chen, G., Guo, C., Faircloth, B.C., et al. (2020). Dense sampling of bird diversity increases power of comparative genomics. Nature 587, 252–257. 10.1038/s41586-020-2873-9.

19. Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., Storz, J.F., Antunes, A., Greenwold, M.J., Meredith, R.W., et al. (2014).

Comparative genomics reveals insights into avian genome evolution and adaptation. Science (80-.). 346, 1311–1320. 10.1126/science.1251385.

20. Craig, R.J., Suh, A., Wang, M., and Ellegren, H. (2018). Natural selection beyond genes: Identification and analyses of evolutionarily conserved elements

21. Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. Pubbl. del R Ist. Super. di Sci. Econ. e Commericiali di Firenze 8, 3–62.

22. O'Brien, T.W. (2002). Evolution of a protein-rich mitochondrial ribosome: implications for human genetic disease. Gene 286, 73–79. 10.1016/S0378-1119(01)00808-3.

23. Hirose, K., Chang, S., Yu, H., Wang, J., Barca, E., Chen, X., Khyeam, S., Tajima, K., Yoneshiro, T., Kajimura, S., et al. (2019). Loss of a novel striated muscle-enriched mitochondrial protein Coq10a enhances postnatal cardiac hypertrophic growth. bioRxiv. 10.1101/755793.

24. Singleton, R.S., Liu-Yi, P., Formenti, F., Ge, W., Sekirnik, R., Fischer, R., Adam, J., Pollard, P.J., Wolf, A., Thalhammer, A., et al. (2014). OGFOD1 catalyzes prolyl hydroxylation of RPS23 and is involved in translation control and stress granule formation. Proc. Natl. Acad. Sci. 111, 4031–4036. 10.1073/pnas.1314482111.

25. Shang, L., Pruett, N.D., and Awgulewitsch, A. (2002). Hoxc12 expression pattern in developing and cycling murine hair follicles. Mech. Dev. 113, 207–210. 10.1016/S0925-4773(02)00022-9.

26. Hostikka, S.L., and Capecchi, M.R. (1998). The mouse Hoxc11 gene: genomic structure and expression pattern. Mech. Dev. 70, 133–145. 10.1016/S0925-4773(97)00182-2.

27. Pasternack, S.M., Refke, M., Paknia, E., Hennies, H.C., Franz, T., Schäfer, N., Fryer, A., van Steensel, M., Sweeney, E., Just, M., et al. (2013). Mutations in SNRPE, which Encodes a Core Protein of the Spliceosome, Cause Autosomal-Dominant Hypotrichosis Simplex. Am. J. Hum. Genet. 92, 81–87. 10.1016/j.ajhg.2012.10.022.

28. Anttonen, A.-K., Laari, A., Kousi, M., Yang, Y.J., Jääskeläinen, T., Somer, M., Siintola, E., Jakkula, E., Muona, M., Tegelberg, S., et al. (2017). ZNHIT3 is defective in PEHO syndrome, a severe encephalopathy with cerebellar granule neuron loss. Brain 140, 1267–1279. 10.1093/brain/awx040.

29. Gouge, J., Satia, K., Guthertz, N., Widya, M., Thompson, A.J., Cousin, P., Dergai, O., Hernandez, N., and Vannini, A. (2015). Redox Signaling by the RNA Polymerase III TFIIB-Related Factor Brf2. Cell 163, 1375–1387. 10.1016/j.cell.2015.11.005.

30. Oppmann, B., Lesley, R., Blom, B., Timans, J.C., Xu, Y., Hunte, B., Vega, F., Yu, N., Wang, J., Singh, K., et al. (2000). Novel p19 Protein Engages IL-12p40 to Form a Cytokine, IL-23, with Biological Activities Similar as Well as Distinct from IL-12. Immunity 13, 715–725. 10.1016/S1074-7613(00)00070-4.

31. Lankford, C.S.R. (2003). A unique role for IL-23 in promoting cellular immunity. J. Leukoc. Biol. 73, 49–56. 10.1189/jlb.0602326.

32. Kopp, T., Lenz, P., Bello-Fernandez, C., Kastelein, R.A., Kupper, T.S., and Stingl, G. (2003). IL-23 Production by Cosecretion of Endogenous p19 and Transgenic p40 in Keratin 14/p40 Transgenic Mice: Evidence for Enhanced Cutaneous Immunity. J. Immunol. 170, 5438–5444. 10.4049/jimmunol.170.11.5438.

33. Mao, H., Pilaz, L.-J., McMahon, J.J., Golzio, C., Wu, D., Shi, L., Katsanis, N., and Silver, D.L. (2015). Rbm8a Haploinsufficiency Disrupts Embryonic Cortical Development Resulting in Microcephaly. J. Neurosci. 35, 7003–7018. 10.1523/JNEUROSCI.0018-15.2015.

34. Lopez, L.M., Mullen, W., Zürbig, P., Harris, S.E., Gow, A.J., Starr, J.M., Porteous, D.J., Mischak, H., and Deary, I.J. (2011). A pilot study of urinary peptides as biomarkers for intelligence in old age. Intelligence 39, 46–53. 10.1016/j.intell.2010.11.001.

35. Balcueva, E.A., Wang, Q., Hughes, H., Kunsch, C., Yu, Z., and Robishaw, J.D. (2000). Human G Protein γ11 and γ14 Subtypes Define a New Functional Subclass. Exp. Cell Res. 257, 310–319. 10.1006/excr.2000.4893.

36. Johnson, D.W., Berg, J.N., Baldwin, M.A., Gallione, C.J., Marondel, I., Yoon, S.-J., Stenzel, T.T., Speer, M., Pericak-Vance, M.A., Diamond, A., et al. (1996). Mutations in the activin receptor–like kinase 1 gene in hereditary haemorrhagic telangiectasia type 2. Nat. Genet. 13, 189–195. 10.1038/ng0696-189.

37. Urness, L.D., Sorensen, L.K., and Li, D.Y. (2000). Arteriovenous malformations in mice lacking activin receptor-like kinase-1. Nat. Genet. 26, 328–331. 10.1038/81634.

38. Seki, T., Yun, J., and Oh, S.P. (2003). Arterial Endothelium-Specific Activin Receptor-Like Kinase 1 Expression Suggests Its Role in Arterialization and Vascular Remodeling. Circ. Res. 93, 682–689. 10.1161/01.RES.0000095246.40391.3B.

39. Lux, A., Salway, F., Dressman, H.K., Kröner-Lux, G., Hafner, M., Day, P.J., Marchuk, D.A., and Garland, J. (2006). ALK1 signalling analysis identifies angiogenesis related genes and reveals disparity between TGF-β and constitutively active receptor induced gene expression. BMC Cardiovasc. Disord. 6, 13. 10.1186/1471-2261-6-13.

40. Wang, X., Gerber, A., Chen, W.-Y., and Roeder, R.G. (2020). Functions of paralogous RNA polymerase III subunits POLR3G and POLR3GL in mouse development. Proc. Natl. Acad. Sci. 117, 15702–15711. 10.1073/pnas.1922821117.

41. James, C., Zhao, T.Y., Rahim, A., Saxena, P., Muthalif, N.A., Uemura, T., Tsuneyoshi, N., Ong, S., Igarashi, K., Lim, C.Y., et al. (2018). MINDY1 Is a Downstream Target of the Polyamines and Promotes Embryonic Stem Cell Self-Renewal. Stem Cells 36, 1170–1178. 10.1002/stem.2830.

42. Seale, L.A., Hashimoto, A.C., Kurokawa, S., Gilman, C.L., Seyedali, A., Bellinger, F.P., Raman, A. V, and Berry, M.J. (2012). Disruption of the Selenocysteine Lyase-Mediated Selenium Recycling Pathway Leads to Metabolic Syndrome in Mice. Mol. Cell. Biol. 32, 4141–4154. 10.1128/MCB.00293-12.

43. Seale, L.A., Gilman, C.L., Hashimoto, A.C., Ogawa-Wong, A.N., and Berry, M.J. (2015). Diet-Induced Obesity in the Selenocysteine Lyase Knockout Mouse. Antioxid. Redox Signal. 23, 761–774. 10.1089/ars.2015.6277.

44. Delpón, E., Cordeiro, J.M., Núñez, L., Thomsen, P.E.B., Guerchicoff, A., Pollevick, G.D., Wu, Y., Kanters, J.K., Larsen, C.T., Hofman-Bang, J., et al. (2008). Functional Effects of KCNE3 Mutation and Its Role in the Development of Brugada Syndrome. Circ. Arrhythmia Electrophysiol. 1, 209–218. 10.1161/CIRCEP.107.748103.

45. Preston, P., Wartosch, L., Günzel, D., Fromm, M., Kongsuphol, P., Ousingsawat, J., Kunzelmann, K., Barhanin, J., Warth, R., and Jentsch, T.J. (2010). Disruption of the K+ Channel β-Subunit KCNE3 Reveals an Important Role in Intestinal and Tracheal Cl− Transport. J. Biol. Chem. 285, 7165–7175. 10.1074/jbc.M109.047829.

46. Guo, Y.-P., Zhi, Y.-R., Liu, T.-T., Wang, Y., and Zhang, Y. (2019). Global Gene Knockout of Kcnip3 Enhances Pain Sensitivity and Exacerbates Negative Emotions in Rats. Front. Mol. Neurosci. 12, 5. 10.3389/fnmol.2019.00005.

47. Deneen, B., Ho, R., Lukaszewicz, A., Hochstim, C.J., Gronostajski, R.M., and Anderson, D.J. (2006). The Transcription Factor NFIA Controls the Onset of Gliogenesis in the Developing Spinal Cord. Neuron 52, 953–968. 10.1016/j.neuron.2006.11.019.

48. Pevny, L.H., and Nicolis, S.K. (2010). Sox2 roles in neural stem cells. Int. J. Biochem. Cell Biol. 42, 421–424. 10.1016/j.biocel.2009.08.018.

49. Uehara, T., Tsuchihashi, T., Yamada, M., Suzuki, H., Takenouchi, T., and Kosaki, K. (2019). CNOT2 haploinsufficiency causes a neurodevelopmental disorder with characteristic facial features. Am. J. Med. Genet. Part A 179, 2506–2509. 10.1002/ajmg.a.61356.

50. Geisler, S., Vollmer, S., Golombek, S., and Kahle, P.J. (2014). UBE2N, UBE2L3 and UBE2D2/3 ubiquitin-conjugating enzymes are essential for parkin-dependent mitophagy. J. Cell Sci. 127, 3280–3293. 10.1242/jcs.146035.

51. Zhu, Q., Yang, G., Chen, B., Liu, F., Li, X., and Liu, L. (2020). Altered Expression of GJD2 Messenger RNA and the Coded Protein Connexin 36 in Negative Lens–induced Myopia of Guinea Pigs. Optom. Vis. Sci. 97, 1080–1088. 10.1097/OPX.0000000000001611.

52. Tsui, D., Vessey, J.P., Tomita, H., Kaplan, D.R., and Miller, F.D. (2013). FoxP2 Regulates Neurogenesis during Embryonic Cortical Development. J. Neurosci. 33, 244–258. 10.1523/JNEUROSCI.1665-12.2013.

53. Zhang, D., Zhao, T., Ang, H.S., Chong, P., Saiki, R., Igarashi, K., Yang, H., and Vardy, L.A. (2012). AMD1 is essential for ESC self-renewal and is translationally down-regulated on differentiation to neural precursor cells. Genes Dev. 26, 461–473. 10.1101/gad.182998.111.

54. Smallwood, P.M., Munoz-Sanjuan, I., Tong, P., Macke, J.P., Hendry, S.H., Gilbert, D.J., Copeland, N.G., Jenkins, N.A., and Nathans, J. (1996). Fibroblast growth factor (FGF) homologous factors: new members of the FGF family implicated in nervous system development. Proc. Natl. Acad. Sci. 93, 9850–9857. 10.1073/pnas.93.18.9850.

55. Lucey, M.M., Wang, Y., Bustin, M., and Duncan, M.K. (2008). Differential expression of the HMGN family of chromatin proteins during ocular development. Gene Expr. Patterns 8, 433–437. 10.1016/j.gep.2008.04.002.

56. Javed, A., Mattar, P., Lu, S., Kruczek, K., Kloc, M., Gonzalez-Cordero, A., Bremner, R., Ali, R.R., and Cayouette, M. (2020). Pou2f1 and Pou2f2 cooperate to control the timing of cone photoreceptor production in the developing mouse retina. Development 147. 10.1242/dev.188730.

57. Tse, E., Smith, A.J.H., Hunt, S., Lavenir, I., Forster, A., Warren, A.J., Grutz, G., Foroni, L., Carlton, M.B.L., Colledge, W.H., et al. (2004). Null Mutation of the Lmo4 Gene or a Combined Null Mutation of the Lmo1 / Lmo3 Genes Causes Perinatal Lethality, and

Lmo4 Controls Neural Tube Development in Mice. Mol. Cell. Biol. 24, 2063–2073. 10.1128/MCB.24.5.2063-2073.2004.

58. Pizzo, L., Jensen, M., Polyak, A., Rosenfeld, J.A., Mannik, K., Krishnan, A., McCready, E., Pichon, O., Le Caignec, C., Van Dijck, A., et al. (2019). Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. Genet. Med. 21, 816–825. 10.1038/s41436-018-0266-3.

59. Malinow, R., Schulman, H., and Tsien, R.W. (1989). Inhibition of Postsynaptic PKC or CaMKII Blocks Induction But Not Expression of LTP. Science (80-.). 245, 862–866. 10.1126/science.2549638.

60. Silva, A.J., Stevens, C.F., Tonegawa, S., and Wang, Y. (1992). Deficient Hippocampal Long-Term Potentiation in α-Calcium-Calmodulin Kinase II Mutant Mice. Science (80-.). 257, 201–206. 10.1126/science.1378648.

61. Hayashi, Y., Shi, S.-H., Esteban, J.A., Piccini, A., Poncer, J.-C., and Malinow, R. (2000). Driving AMPA Receptors into Synapses by LTP and CaMKII: Requirement for GluR1 and PDZ Domain Interaction. Science (80-.). 287, 2262–2267. 10.1126/science.287.5461.2262.

62. Benke, T.A., Lüthi, A., Isaac, J.T.R., and Collingridge, G.L. (1998). Modulation of AMPA receptor unitary conductance by synaptic activity. Nature 393, 793–797. 10.1038/31709.

63. Derkach, V., Barria, A., and Soderling, T.R. (1999). Ca 2+ /calmodulin-kinase II enhances channel conductance of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionate type glutamate receptors. Proc. Natl. Acad. Sci. 96, 3269–3274. 10.1073/pnas.96.6.3269.

64. Giese, K.P., and Mizuno, K. (2013). The roles of protein kinases in learning and memory. Learn. Mem. 20, 540–552. 10.1101/lm.028449.112.

65. Buard, I., Coultrap, S.J., Freund, R.K., Lee, Y.-S., Dell'Acqua, M.L., Silva, A.J., and Bayer, K.U. (2010). CaMKII "Autonomy" Is Required for Initiating But Not for Maintaining Neuronal Long-Term Information Storage. J. Neurosci. 30, 8214–8220. 10.1523/JNEUROSCI.1469-10.2010.

66. Vigil, F.A., Mizuno, K., Lucchesi, W., Valls-Comamala, V., and Giese, K.P. (2017). Prevention of long-term memory loss after retrieval by an endogenous CaMKII inhibitor. Sci. Rep. 7, 4040. 10.1038/s41598-017-04355-8.

67. O'Rourke, T., Martins, P.T., Asano, R., Tachibana, R.O., Okanoya, K., and Boeckx, C. (2021). Capturing the Effects of Domestication on Vocal Learning Complexity. Trends Cogn. Sci. 25, 722. 10.1016/j.tics.2021.05.002.

68. SATTERLEE, D.G., and JOHNSON, W.A. (1988). Selection of Japanese Quail for Contrasting Blood Corticosterone Response to Immobilization. Poult. Sci. 67, 25–32. 10.3382/ps.0670025.

69. Khatri, B., Kang, S., Shouse, S., Anthony, N., Kuenzel, W., and Kong, B.C. (2019). Copy number variation study in Japanese quail associated with stress related traits using whole genome re-sequencing data. PLoS One 14, e0214543. 10.1371/journal.pone.0214543.

70. Lai, C.S.L., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413, 519–523. 10.1038/35097076.

71. Fisher, S.E., and Scharff, C. (2009). FOXP2 as a molecular window into speech and language. Trends Genet. 25, 166–177. 10.1016/j.tig.2009.03.002.

72. Bacon, C., and Rappold, G.A. (2012). The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders. Hum. Genet. 131, 1687–1698. 10.1007/s00439-012-1193-z.

73. Sollis, E., Graham, S.A., Vino, A., Froehlich, H., Vreeburg, M., Dimitropoulou, D., Gilissen, C., Pfundt, R., Rappold, G.A., Brunner, H.G., et al. (2016). Identification and functional characterization of de novo FOXP1 variants provides novel insights into the etiology of neurodevelopmental disorder. Hum. Mol. Genet. 25, 546–557. 10.1093/hmg/ddv495.

74. Siper, P.M., De Rubeis, S., Trelles, M.D.P., Durkin, A., Di Marino, D., Muratet, F., Frank, Y., Lozano, R., Eichler, E.E., Kelly, M., et al. (2017). Prospective investigation of FOXP1 syndrome. Mol. Autism 8, 57. 10.1186/s13229-017-0172-6.

75. Morgan, A.T., and Webster, R. (2018). Aetiology of childhood apraxia of speech: A clinical practice update for paediatricians. J. Paediatr. Child Health 54, 1090–1095. 10.1111/jpc.14150.

76. Teramitsu, I. (2004). Parallel FoxP1 and FoxP2 Expression in Songbird and Human Brain Predicts Functional Interaction. J. Neurosci. 24, 3152–3163. 10.1523/JNEUROSCI.5589-03.2004.

77. Mendoza, E., Tokarev, K., Düring, D.N., Retamosa, E.C., Weiss, M., Arpenik, N., and Scharff, C. (2015). Differential coexpression of FoxP1, FoxP2, and FoxP4 in the Zebra Finch (Taeniopygia guttata) song system. J. Comp. Neurol. 523, 1318–1340. 10.1002/cne.23731.

78. Norton, P., Barschke, P., Scharff, C., and Mendoza, E. (2019). Differential Song Deficits after Lentivirus-Mediated Knockdown of FoxP1, FoxP2, or FoxP4 in Area X of Juvenile Zebra Finches. J. Neurosci. 39, 9782–9796. 10.1523/JNEUROSCI.1250-19.2019.

79. Garcia-Oscos, F., Koch, T., Pancholi, H., Trusel, M., Daliparthi, V., Ayhan, F., Co, M., Alam, D.H., Holdway, J.E., Konopka, G., et al. (2021). Autism-linked gene FoxP1 selectively regulates the cultural transmission of learned vocalizations. Sci. Adv. 7. 10.1101/2020.03.14.992016.

80. Oikkonen, J., Huang, Y., Onkamo, P., Ukkola-Vuoti, L., Raijas, P., Karma, K., Vieland, V.J., and Järvelä, I. (2015). A genome-wide linkage and association study of musical aptitude identifies loci containing genes related to inner ear development and neurocognitive functions. Mol. Psychiatry 20, 275–282. 10.1038/mp.2014.8.

81. Park, H., Lee, S., Kim, H.-J., Ju, Y.S., Shin, J.-Y., Hong, D., von Grotthuss, M., Lee, D.-S., Park, C., Kim, J.H., et al. (2012). Comprehensive genomic analyses associate UGT8 variants with musical ability in a Mongolian population. J. Med. Genet. 49, 747–752. 10.1136/jmedgenet-2012-101209.

82. Oikkonen, J., Kuusi, T., Peltonen, P., Raijas, P., Ukkola-Vuoti, L., Karma, K., Onkamo, P., and Järvelä, I. (2016). Creative Activities in Music – A Genome-Wide Linkage Analysis. PLoS One 11, e0148679. 10.1371/journal.pone.0148679.

83. Adelinik, and Papian Curcumin modulates the level of oxidative status through the up-regulation of follicle-stimulating hormone in the testes of lead-exposed mice. J. Chem. Pharm. Res.

84. Zambonelli, P., Gaffo, E., Zappaterra, M., Bortoluzzi, S., and Davoli, R. (2016). Transcriptional profiling of subcutaneous adipose tissue in Italian Large White pigs divergent for backfat thickness. Anim. Genet. 47, 306–323. 10.1111/age.12413.

85. Di Natale, M.R., Soch, A., Ziko, I., De Luca, S.N., Spencer, S.J., and Sominsky, L. (2019). Chronic predator stress in female mice reduces primordial follicle numbers: implications for the role of ghrelin. J. Endocrinol. 241, 201–219. 10.1530/JOE-19-0109.

86. Kim, Y.-R., Jacobs, J.S., Li, Q., Gaddam, R.R., Vikram, A., Liu, J., Kassan, M., Irani, K., and Kumar, S. (2019). SUMO2 regulates vascular endothelial function and oxidative stress in mice. Am. J. Physiol. Circ. Physiol. 317, H1292–H1300. 10.1152/ajpheart.00530.2019.

87. Datwyler, A.L., Lättig-Tünnemann, G., Yang, W., Paschen, W., Lee, S.L.L., Dirnagl, U., Endres, M., and Harms, C. (2011). SUMO2/3 Conjugation is an Endogenous Neuroprotective Mechanism. J. Cereb. Blood Flow Metab. 31, 2152–2159. 10.1038/jcbfm.2011.112.

88. Papale, L.A., Madrid, A., Li, S., and Alisch, R.S. (2017). Early-life stress links 5-hydroxymethylcytosine to anxiety-related behaviors. Epigenetics 1

89. Boehler, N.A., Fung, S.W., Hegazi, S., Cheng, A.H., and Cheng, H.-Y.M. (2021). Sox2 Ablation in the Suprachiasmatic Nucleus Perturbs Anxiety- and Depressive-like Behaviors. Neurol. Int. 13, 541–554. 10.3390/neurolint13040054.

90. Farahmandfar, M., Akbarabadi, A., Bakhtazad, A., and Zarrindast, M.-R. (2017). Recovery from ketamine-induced amnesia by blockade of GABA-A receptor in the medial prefrontal cortex of mice. Neuroscience 344, 48–55. 10.1016/j.neuroscience.2016.02.056.

91. Ito, K., Inoue, T., Yokoyama, K., Morita, M., Suzuki, T., and Yamamoto, T. (2011). CNOT2 depletion disrupts and inhibits the CCR4-NOT deadenylase complex and induces apoptotic cell death. Genes to Cells 16, 368–379. 10.1111/j.1365-2443.2011.01492.x.

92. Kim, E.Y., Zhang, Y., Ye, B., Segura, A.M., Beketaev, I., Xi, Y., Yu, W., Chang, J., Li, F., and Wang, J. (2015). Involvement of activated SUMO-2 conjugation in cardiomyopathy. Biochim. Biophys. Acta - Mol. Basis Dis. 1852, 1388–1399. 10.1016/j.bbadis.2015.03.013.

93. Zhou, Z., Chen, Y., Ni, W., and Liu, T. (2019). Upregulation of Nuclear Factor IA Suppresses Oxidized Low-Density Lipoprotein-Induced Endoplasmic Reticulum Stress and Apoptosis in Human Umbilical Vein Endothelial Cells. Med. Sci. Monit. 25, 1009–1016. 10.12659/MSM.912132.

94. Peñaranda-Fajardo, N.M., Meijer, C., Liang, Y., Dijkstra, B.M., Aguirre-Gamboa, R., den Dunnen, W.F.A., and Kruyt, F.A.E. (2019). ER stress and UPR activation in glioblastoma: identification of a noncanonical PERK mechanism regulating GBM stem cells through SOX2 modulation. Cell Death Dis. 10, 690. 10.1038/s41419-019-1934-1.

95. Zar, J.H. (1972). Significance Testing of the Spearman Rank Correlation Coefficient. J. Am. Stat. Assoc. 67, 578–580. 10.1080/01621459.1972.10481251.

96. Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C.F., Olason, P., and Ellegren, H. (2014). 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. 10.1111/mec.12810.

97. Singhal, S., Leffler, E.M., Sannareddy, K., Turner, I., Venn, O., Hooper, D.M., Strand, A.I., Li, Q., Raney, B., Balakrishnan, C.N., et al. (2015). Stable recombination hotspots in birds. Science (80-.). 350, 928–932. 10.1126/science.aad0843.

98. Kawakami, T., Mugal, C.F., Suh, A., Nater, A., Burri, R., Smeds, L., and Ellegren, H. (2017). Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. Mol. Ecol. 38, 42–49. 10.1111/mec.14197.

99. Kim, Y., and Nielsen, R. (2004). Linkage Disequilibrium as a Signature of Selective Sweeps. Genetics 167, 1513–1524. 10.1534/genetics.103.025387.

100. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A Map of Recent Positive Selection in the Human Genome. PLoS Biol. 4, e72. 10.1371/journal.pbio.0040072.

101. Ennis, S. (2007). Linkage Disequilibrium as a Tool for Detecting Signatures of Natural Selection. In Methods Mol. Biol., pp. 59–70. 10.1007/978-1-59745-389-9_5.

102. Coop, G., Wen, X., Ober, C., Pritchard, J.K., and Przeworski, M. (2008). High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. Science (80-.). 319, 1395–1398. 10.1126/science.1151851.

103. Zaitlen, N., Huntsman, S., Hu, D., Spear, M., Eng, C., Oh, S.S., White, M.J., Mak, A., Davis, A., Meade, K., et al. (2017). The Effects of Migration and Assortative Mating on Admixture Linkage Disequilibrium. Genetics 205, 375–383. 10.1534/genetics.116.192138.

104. Bürger, R., and Akerman, A. (2011). The effects of linkage and gene flow on local adaptation: A two-locus continent–island model. Theor. Popul. Biol. 80, 272–288. 10.1016/j.tpb.2011.07.002.

105. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The Structure of Haplotype Blocks in the Human Genome. Science (80-.). 296, 2225–2229. 10.1126/science.1069424.

106. Zhou, Y., Qiu, H., and Xu, S. (2017). Modeling Continuous Admixture Using Admixture-Induced Linkage Disequilibrium. Sci. Rep. 7, 43054. 10.1038/srep43054.

107. Smith, C.C.R., Flaxman, S.M., Scordato, E.S.C., Kane, N.C., Hund, A.K., Sheta, B.M., and Safran, R.J. (2018). Demographic inference in barn swallows using whole-genome data shows signal for bottleneck and subspecies differentiation during the Holocene. Mol. Ecol. 27, 4200–4212. 10.1111/mec.14854.

108. Monteggia, L.M., Barrot, M., Powell, C.M., Berton, O., Galanis, V., Gemelli, T., Meuth, S., Nagy, A., Greene, R.W., and Nestler, E.J. (2004). Essential role of brain-derived neurotrophic factor in adult hippocampal function. Proc. Natl. Acad. Sci. 101, 10827–10832. 10.1073/pnas.0402141101.

109. Bramham, C.R., and Messaoudi, E. (2005). BDNF function in adult synaptic plasticity: The synaptic consolidation hypothesis. Prog. Neurobiol. 76, 99–125. 10.1016/j.pneurobio.2005.06.003.

110. Li, X.-C., Jarvis, E.D., Alvarez-Borda, B., Lim, D.A., and Nottebohm, F. (2000). A relationship between behavior, neurotrophin expression, and new neuron survival. Proc. Natl. Acad. Sci. 97, 8584–8589. 10.1073/pnas.140222497.

111. Dittrich, F., Feng, Y., Metzdorf, R., and Gahr, M. (1999). Estrogen-inducible, sex-specific expression of brain-derived neurotrophic factor mRNA in a forebrain song control nucleus of the juvenile zebra finch. Proc. Natl. Acad. Sci. 96, 8241–8246. 10.1073/pnas.96.14.8241.

112. Dittrich, F., ter Maat, A., Jansen, R.F., Pieneman, A., Hertel, M., Frankl-Vilches, C., and Gahr, M. (2013). Maximized song learning of juvenile male zebra finches following BDNF expression in the HVC. Eur. J. Neurosci. 38, 3338–3344. 10.1111/ejn.12329.

113. Wang, H., Sawai, A., Toji, N., Sugioka, R., Shibata, Y., Suzuki, Y., Ji, Y., Hayase, S., Akama, S., Sese, J., et al. (2019). Transcriptional regulatory divergence underpinning species-specific learned vocalization in songbirds. PLOS Biol. 17, e3000476. 10.1371/journal.pbio.3000476.

114. Sieber-Blum, M. (1991). Role of the neurotrophic factors BDNF and NGF in the commitment of pluripotent neural crest cells. Neuron 6, 949–955. 10.1016/0896-6273(91)90235-R.

115. Wilkins, A.S., Wrangham, R.W., and Fitch, W.T. (2014). The “Domestication Syndrome” in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics. Genetics 197, 795–808. 10.1534/genetics.114.165423.

116. Sánchez-Villagra, M.R., Geiger, M., and Schneider, R.A. (2016). The taming of the neural crest: a developmental perspective on the origins of morphological covariation in domesticated mammals. R. Soc. Open Sci. 3, 160107. 10.1098/rsos.160107.

117. Notaras, M., and van den Buuse, M. (2020). Neurobiology of BDNF in fear memory, sensitivity to stress, and stress-related disorders. Mol. Psychiatry 25, 2251–2274. 10.1038/s41380-019-0639-2.

118. Maynard, K.R., Hill, J.L., Calcaterra, N.E., Palko, M.E., Kardian, A., Paredes, D., Sukumar, M., Adler, B.D., Jimenez, D. V, Schloesser, R.J., et al. (2016). Functional Role of BDNF Production from Unique Promoters in Aggression and Serotonin Signaling. Neuropsychopharmacology 41, 1943–1955. 10.1038/npp.2015.349.

119. Yossifoff, M., Kisliouk, T., and Meiri, N. (2008). Dynamic changes in DNA methylation during thermal control establishment affect CREB binding to the brain-derived neurotrophic factor promoter. Eur. J. Neurosci. 28, 2267–2277. 10.1111/j.1460-9568.2008.06532.x.

120. George, J.M., Bell, Z.W., Condliffe, D., Dohrer, K., Abaurrea, T., Spencer, K., Leitão, A., Gahr, M., Hurd, P.J., and Clayton, D.F. (2020). Acute social isolation alters neurogenomic state in songbird forebrain. Proc. Natl. Acad. Sci. 117, 23311–23316. 10.1073/pnas.1820841116.

121. Lubin, F.D., Roth, T.L., and Sweatt, J.D. (2008). Epigenetic Regulation of bdnf Gene Transcription in the Consolidation of Fear Memory. J. Neurosci. 28, 10576–10586. 10.1523/JNEUROSCI.1786-08.2008.

122. de Lau, W., Peng, W.C., Gros, P., and Clevers, H. (2014). The R-spondin/Lgr5/Rnf43 module: regulator of Wnt signal strength. Genes Dev. 28, 305–316. 10.1101/gad.235473.113.

123. Li, Z., Zhang, W., and Mulholland, M.W. (2015). LGR4 and Its Role in Intestinal Protection and Energy Metabolism. Front. Endocrinol. (Lausanne). 6, 131. 10.3389/fendo.2015.00131.

124. Huang, X., Otecko, N.O., Peng, M., Weng, Z., Li, W., Chen, J., Zhong, M., Zhong, F., Jin, S., Geng, Z., et al. (2020). Genome-wide genetic structure and selection signatures for color in 10 traditional Chinese yellow-feathered chicken breeds. BMC Genomics 21, 316. 10.1186/s12864-020-6736-4.

125. Weng, Z., Xu, Y., Zhong, M., Li, W., Chen, J., Zhong, F., Du, B., Zhang, B., and Huang, X. (2022). Runs of homozygosity analysis reveals population characteristics of yellow-feathered chickens using re-sequencing data. Br. Poult. Sci. 63, 307–315. 10.1080/00071668.2021.2003752.

126. Hou, Q., Han, S., Yang, L., Chen, S., Chen, J., Ma, N., Wang, C., Tang, J., Chen, X., Chen, F., et al. (2019). The Interplay of MicroRNA-34a, LGR4, EMT-Associated Factors, and MMP2 in Regulating Uveal Melanoma Cells. Investig. Opthalmology Vis. Sci. 60, 4503. 10.1167/iovs.18-26477.

127. Ren, X., Xia, W., Xu, P., Shen, H., Dai, X., Liu, M., Shi, Y., Ye, X., and Dang, Y. (2020). Lgr4 Deletion Delays the Hair Cycle and Inhibits the Activation of Hair Follicle Stem Cells. J. Invest. Dermatol. 140, 1706-1712.e4. 10.1016/j.jid.2019.12.034.

128. Yuan, J., Li, S., Sheng, Z., Zhang, M., Liu, X., Yuan, Z., Yang, N., and Chen, J. (2022). Genome-wide run of homozygosity analysis reveals candidate genomic regions associated with environmental adaptations of Tibetan native chickens. BMC Genomics 23, 91. 10.1186/s12864-021-08280-z.

129. O'Reilly, P.F., Birney, E., and Balding, D.J. (2008). Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res. 18, 1304–1313. 10.1101/gr.067181.107.

130. Dierckxsens, N., Mardulyn, P., and Smits, G. (2016). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 45, gkw955. 10.1093/nar/gkw955.

131. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. 29, 24–26. 10.1038/nbt.1754.