Howard Hughes Medical Institute

Published as: *Cell Rep*. 2023 January 31; 42(1): 111992–111992.

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,*], **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

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

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.

SUPPLEMENTAL INFORMATION

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

DECLARATION OF INTERESTS

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

[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

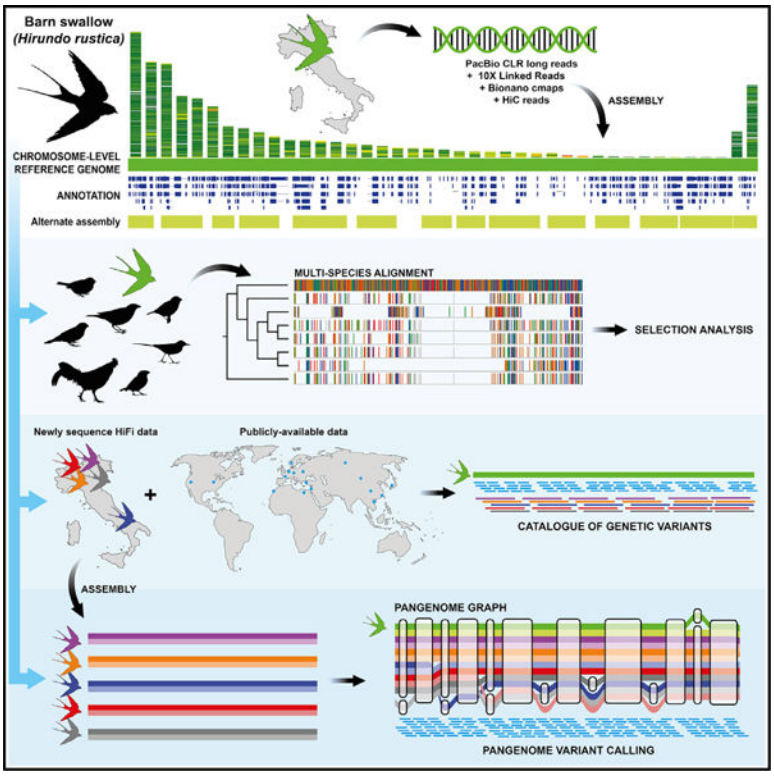
[17]These authors contributed equally

[18]Lead contact

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.

Graphical abstract



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.

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

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 $\rho = 0.83$, n = 108,010, $p < 2.2 \times 10^{-16}$). 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 $\rho = 0.35$, n = 38, $p < 3.4 \times 10^{-2}$) 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 $\rho = -0.80$, n = 38, $p < 9.5 \times 10^{-8}$), with microchromosomes richer in accelerated sites than other chromosome types (Wilcoxon test, W = 50, $p = 4.6 \times 10^{-5}$), 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] ($\chi^2 = 2.03 \times 10^7$, df = 1, $p < 2 \times 10^{-16}$). 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 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 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. 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

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 mg/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, after4h 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 −binaries-Mode 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-extrachr –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 R1 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 S9D).

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–S3E). 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–S3I). 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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.

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 CCR, Flaxman SM, Scordato ESC, Kane NC, Hund AK, Sheta BM, 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. 10.1111/mec.14854. [PubMed: 30176075]
3. Lombardo G, Rambaldi Migliore N, Colombo G, Capodiferro MR, Formenti G, Caprioli M, et al. (2022). The mitogenome relationships and phylogeography of barn swallows (*Hirundo rustica*). Mol. Biol. Evol 39, msac113. 10.1093/molbev/msac113. [PubMed: 35617136]
4. Johnston RF (2001). Synanthropic birds of North America. In Avian Ecology and Conservation in an Urbanizing World, Marzluff JM, Bowman R, and Donnelly R, eds. (Springer US), pp. 49–67. 10.1007/978-1-4615-1531-9_3.
5. Krajcarz M, Krajcarz MT, Baca M, Baumann C, Van Neer W, Popović D, Sudo1-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. 10.1073/pnas.1918884117. [PubMed: 32661161]
6. Turner A (2010). The Barn Swallow (Poyser) 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. 10.1038/srep45412. [PubMed: 28361883]
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. 10.1007/s00442-017-3918-0. [PubMed: 28741127]
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. 10.1371/journal.pone.0096428. [PubMed: 24826890]
10. Pap PL, Osváth G, Aparicio JM, 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. 10.1371/journal.pone.0130844. [PubMed: 26110255]

11. Pap PL, Fülö p A, Adamkova M, Cepak J, Michalkova R, Safran RJ, Stermin AN, Tomasek O, Vágási CI, 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. 10.1002/ece3.5629. [PubMed: 31641472]
12. Rhie A, McCarthy SA, 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. 10.1038/s41586-021-03451-0. [PubMed: 33911273]
13. Garg V, Dudchenko O, Wang J, Khan AW, Gupta S, Kaur P, Han K, Saxena RK, Kale SM, 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. 10.1016/j.jare.2021.10.009. [PubMed: 36513421]
14. Safran RJ, Scordato ESC, Wilkins MR, Hubbard JK, Jenkins BR, Albrecht T, Flaxman SM, 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. 10.1111/mec.13740. [PubMed: 27357267]
15. Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC, et al. (2020). Dense sampling of bird diversity increases power of comparative genomics. Nature 587, 252–257. 10.1038/s41586-020-2873-9. [PubMed: 33177665]
16. Formenti G, Chiara M, Poveda L, Francoijs K-J, Bonisoli-Alquati A, Canova L, Gianfranceschi L, Horner DS, 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. 10.1093/gigascience/giy142. [PubMed: 30496513]
17. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, and DePristo MA (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genom 13, 375. 10.1186/1471-2164-13-375.
18. Howe K, and Wood JMD (2015). Using optical mapping data for the improvement of vertebrate genome assemblies. Gigascience 4, 10. 10.1186/s13742-015-0052-y. [PubMed: 25789164]
19. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. (2022). A draft human pangenome reference. Preprint at bioRxiv 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. 10.1016/j.tig.2022.06.011. [PubMed: 35817620]
21. Ranallo-Benavidez TR, Jaron KS, and Schatz MC (2020). Genome-Scope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun 11, 1432. 10.1038/s41467-020-14998-3. [PubMed: 32188846]
22. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, and Zdobnov EM (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol 35, 543–548. 10.1093/molbev/msx319. [PubMed: 29220515]
23. Tomaszkiewicz M, Medvedev P, and Makova KD (2017). Y and W Chromosome assemblies: approaches and discoveries. Trends Genet 33, 266–282. 10.1016/j.tig.2017.01.008. [PubMed: 28236503]
24. Malinovskaya LP, Tishakova K, Shnaider EP, Borodin PM, and Torgasheva AA (2020). Heterochiasmy and sexual dimorphism: the case of the barn swallow (*Hirundo rustica, Hirundinidae, aves*). Genes 11, 1119. 10.3390/genes11101119. [PubMed: 32987748]
25. Rhie A, Walenz BP, Koren S, and Phillippy AM (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21, 245. 10.1186/s13059-020-02134-9. [PubMed: 32928274]
26. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, and Zdobnov EM (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212. 10.1093/bioinformatics/btv351. [PubMed: 26059717]

27. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42, D756–D763. 10.1093/nar/gkt1114. [PubMed: 24259432]
28. Francis WR, and Wörheide G (2017). Similar ratios of introns to intergenic sequence across animal genomes. Genome Biol. Evol 9, 1582–1598. 10.1093/gbe/evx103. [PubMed: 28633296]
29. Burt DW (2002). Origin and evolution of avian microchromosomes. Cytogenet. Genome Res 96, 97–112. 10.1159/000063018. [PubMed: 12438785]
30. Kim J, Lee C, Ko BJ, Yoo DA, 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 10.1101/2021.04.09.438906.
31. Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, and Jarvis ED (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. 10.1093/gigascience/gix085.
32. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, 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. 10.1038/s41586-020-2871-y. [PubMed: 33177663]
33. Hubisz MJ, Pollard KS, and Siepel A (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. Brief. Bioinform 12, 41–51. 10.1093/bib/bbq072. [PubMed: 21278375]
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. 10.1111/j.2517-6161.1995.tb02031.x.
35. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15, 1034–1050. 10.1101/gr.3715005. [PubMed: 16024819]
36. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346, 1311–1320. 10.1126/science.1251385. [PubMed: 25504712]
37. Craig RJ, 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. 10.1111/mec.14462. [PubMed: 29226517]
38. Axelsson E, Webster MT, Smith NGC, Burt DW, 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. 10.1101/gr.3021305. [PubMed: 15590944]
39. Schield DR, Scordato ESC, Smith CCR, Carter JK, Cherkaoui SI, Gombobaatar S, Hajib S, Hanane S, Hund AK, Koyama K, et al. (2021). Sex-linked genetic diversity and differentiation in a globally distributed avian species complex. Mol. Ecol 30, 2313–2332. 10.1111/mec.15885. [PubMed: 33720472]
40. von Rönn JAC, Shafer ABA, Wolf JBW, Hybridization GOF, von Rönn JAC, Shafer ABA, Wolf JBW, and Hybridization GOF (2016). Disruptive selection without genome-wide evolution across a migratory divide. Mol. Ecol 25, 2529–2541. 10.1111/mec.13521. [PubMed: 26749140]
41. Scordato ESC, Wilkins MR, Semenov G, Rubtsov AS, Kane NC, and Safran RJ (2017). Genomic variation across two barn swallow hybrid zones reveals traits associated with divergence in sympatry and allopatry. Mol. Ecol 26, 5676–5691. 10.1111/mec.14276. [PubMed: 28777875]
42. Smeds L, Warmuth V, Bolivar P, Uebbing S, Burri R, Suh A, Nater A, Burěs S, Garamszegi LZ, Hogner S, et al. (2015). Evolutionary analysis of the female-specific avian W chromosome. Nat. Commun 6, 7330. 10.1038/ncomms8330. [PubMed: 26040272]
43. Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL, et al. (2017). Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science 358, 951–954. 10.1126/science.aao0960. [PubMed: 29146814]
44. Corcoran P, Gossmann TI, Barton HJ, 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. 10.1093/gbe/evx213. [PubMed: 29045655]

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. 10.1186/s12711-021-00657-8. [PubMed: 34301193]
46. Slatkin M (2008). Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet 9, 477–485. 10.1038/nrg2361. [PubMed: 18427557]
47. Joiret M, Mahachie John JM, Gusareva ES, and Van Steen K (2019). Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. BioData Min 12, 11. 10.1186/s13040-019-0199-7. [PubMed: 31198442]
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. 10.1007/s13258-017-0539-2. [PubMed: 28706593]
49. Pritchard JK, Stephens M, Rosenberg NA, and Donnelly P (2000). Association mapping in structured populations. Am. J. Hum. Genet 67, 170–181. 10.1086/302959. [PubMed: 10827107]
50. Stapley J, Birkhead TR, Burke T, and Slate J (2010). Pronounced inter- and intrachromosomal variation in linkage disequilibrium across the zebra finch genome. Genome Res 20, 496–502. 10.1101/gr.102095.109. [PubMed: 20357051]
51. Kapusta A, and Suh A (2017). Evolution of bird genomes-a transposon's-eye view. Ann. N. Y. Acad. Sci 1389, 164–185. 10.1111/nyas.13295. [PubMed: 27997700]
52. Cheng H, Concepcion GT, Feng X, Zhang H, and Li H (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175. 10.1038/s41592-020-01056-5. [PubMed: 33526886]
53. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. Brief. Bioinform 19, 118–135. 10.1093/bib/bbw089.
54. Sherman RM, and Salzberg SL (2020). Pan-genomics in the human genome era. Nat. Rev. Genet 21, 243–254. 10.1038/s41576-020-0210-7. [PubMed: 32034321]
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. 10.1073/pnas.88.2.507. [PubMed: 1988950]
56. Samadashwily GM, Dayn A, and Mirkin SM (1993). Suicidal nucleotide sequences for DNA polymerization. EMBO J. 12, 4975–4983. 10.1002/j.1460-2075.1993.tb06191.x. [PubMed: 8262040]
57. Mirkin SM, and Frank-Kamenetskii MD (1994). H-DNA and related structures. Annu. Rev. Biophys. Biomol. Struct 23, 541–576. 10.1146/annurev.bb.23.060194.002545. [PubMed: 7919793]
58. Sirén J, Garrison E, Novak AM, Paten B, and Durbin R (2020). Haplotype-aware graph indexes. Bioinformatics 36, 400–407. 10.1093/bioinformatics/btz575. [PubMed: 31406990]
59. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol 36, 875–879. 10.1038/nbt.4227. [PubMed: 30125266]
60. Papale LA, Madrid A, Li S, and Alisch RS (2017). Early-life stress links 5-hydroxymethylcytosine to anxiety-related behaviors. Epigenetics 12, 264–276. 10.1080/15592294.2017.1285986. [PubMed: 28128679]
61. Vigil FA, Mizuno K, Lucchesi W, Valls-Comamala V, and Giese KP (2017). Prevention of long-term memory loss after retrieval by an endogenous CaMKII inhibitor. Sci. Rep 7, 4040. 10.1038/s41598-017-04355-8. [PubMed: 28642476]
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. 10.1038/s41380-019-0639-2. [PubMed: 31900428]
63. O'Rourke T, Martins PT, Asano R, Tachibana RO, Okanoya K, and Boeckx C (2021). Capturing the effects of domestication on vocal learning complexity. Trends Cogn. Sci 25, 462–474. 10.1016/j.tics.2021.05.002. [PubMed: 33810982]
64. Boehler NA, Fung SW, Hegazi S, Cheng AH, and Cheng H-YM (2021). Sox2 ablation in the suprachiasmatic nucleus perturbs anxiety- and depressive-like behaviors. Neurol. Int 13, 541–554. 10.3390/neurolint13040054. [PubMed: 34842772]

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 song-birds. PLoS Biol 17, e3000476. 10.1371/journal.pbio.3000476. [PubMed: 31721761]

66. Kawakami T, Mugal CF, 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. 10.1111/mec.14197. [PubMed: 28597534]

67. O'Reilly PF, Birney E, and Balding DJ (2008). Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res 18, 1304–1313. 10.1101/gr.067181.107. [PubMed: 18617692]

68. Langmead B, and Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. 10.1038/nmeth.1923. [PubMed: 22388286]

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. 10.1093/nar/gkw955. [PubMed: 28204566]

70. Donath A, Jühling F, Al-Arab M, Bernhart SH, Reinhardt F, Stadler PF, Middendorf M, and Bernt M (2019). Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. Nucleic Acids Res 47, 10543–10552. 10.1093/nar/gkz833. [PubMed: 31584075]

71. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, and Phillippy AM (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17, 132. 10.1186/s13059-016-0997-x. [PubMed: 27323842]

72. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10, 563–569. 10.1038/nmeth.2474. [PubMed: 23644548]

73. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, 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. [PubMed: 27749838]

74. Guan D, McCarthy SA, 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. [PubMed: 31971576]

75. Li H, and Durbin R (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595. 10.1093/bioinformatics/btp698. [PubMed: 20080505]

76. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, and Koren S (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput. Biol 15, e1007273. 10.1371/journal.pcbi.1007273. [PubMed: 31433799]

77. Garrison E, and Marth G (2012). Haplotype-based variant detection from short-read sequencing. Preprint at arXiv 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. 10.1093/bioinformatics/btp352. [PubMed: 19505943]

79. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience 10, giab008. 10.1093/gigascience/giab008. [PubMed: 33590861]

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. 10.1093/bioinformatics/btw159. [PubMed: 27153597]

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

82. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, and Salzberg SL (2004). Versatile and open software for comparing large genomes. Genome Biol 5, R12. 10.1186/gb-2004-5-2-r12. [PubMed: 14759262]

83. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, and Carey VJ (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol 9, e1003118. 10.1371/journal.pcbi.1003118. [PubMed: 23950696]

84. Pedersen BS, and Quinlan AR (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34, 867–868. 10.1093/bioinformatics/btx699. [PubMed: 29096012]

85. Morgulis A, Gertz EM, Schäffer AA, and Agarwala R (2006). WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22, 134–141. 10.1093/bioinformatics/bti774. [PubMed: 16287941]

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. 10.1002/0471250953.bi0410s25.

87. Quinlan AR, and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. 10.1093/bioinformatics/btq033. [PubMed: 20110278]

88. Kumar S, Stecher G, Suleski M, and Hedges SB (2017). TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol 34, 1812–1819. 10.1093/molbev/msx116. [PubMed: 28387841]

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. 10.1093/bioinformatics/btt128. [PubMed: 23505295]

90. Luo W, Friedman MS, Shedden K, Hankenson KD, and Woolf PJ (2009). GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinf 10, 161. 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. 10.1093/bioinformatics/bti525. [PubMed: 16082012]

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. 10.1093/molbev/msy096. [PubMed: 29722887]

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. 10.1093/bioinformatics/btw354. [PubMed: 27312411]

94. Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J 17, 10–12. 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 CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156–2158. 10.1093/bioinformatics/btr330. [PubMed: 21653522]

97. Thorvaldsdóttir H, Robinson JT, and Mesirov JP (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform 14, 178–192. 10.1093/bib/bbs017. [PubMed: 22517427]

98. Gel B, and Serra E (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33, 3088–3090. 10.1093/bioinformatics/btx346. [PubMed: 28575171]

99. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet 81, 559–575. 10.1086/519795. [PubMed: 17701901]

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. 10.1093/bib/bbaa227. [PubMed: 33126247]

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. 10.2174/1574893611666160907111325.

102. Martin M, Patterson M, Garg S, O Fischer S, Pisanti N, Klau GW, Schöenhuth A, and Marschall T (2016). WhatsHap: fast and accurate read-based phasing. Preprint at bioRxiv 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. 10.1093/bioinformatics/bts115. [PubMed: 22402612]
104. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol 36, 983–987. 10.1038/nbt.4235. [PubMed: 30247488]
105. Yun T, Li H, Chang P-C, Lin MF, Carroll A, and McLean CY (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics 36, 5582–5589. 10.1093/bioinformatics/btaa1081. [PubMed: 33399819]
106. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol 37, 1155–1162. 10.1038/s41587-019-0217-9. [PubMed: 31406327]
107. Hall M (2022). Rasusa: randomly subsample sequencing reads to a specified coverage. J. Open Source Softw 7, 3941. 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. 10.1186/s13059-020-02168-z. [PubMed: 33066802]
109. Zhang X, Kaplow IM, Wirthlin M, Park TY, and Pfenning AR (2020). HALPER facilitates the identification of regulatory element orthologs across species. Bioinformatics 36, 4339–4340. 10.1093/bioinformatics/btaa493. [PubMed: 32407523]
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. 10.1093/bioinformatics/btu393. [PubMed: 24930139]
112. Gu Z, Eils R, and Schlesner M (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849. 10.1093/bioinformatics/btw313. [PubMed: 27207943]
113. Beyer W, Novak AM, Hickey G, Chan J, Tan V, Paten B, and Zerbino DR (2019). Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics 35, 5318–5320. 10.1093/bioinformatics/btz597. [PubMed: 31368484]
114. Challis R, Richards E, Rajan J, Cochrane G, and Blaxter M (2020). BlobToolKit–Interactive quality assessment of genome assemblies. G3 10, 1361–1374. 10.1534/g3.119.400908. [PubMed: 32071071]
115. Cabanettes F, and Klopp C (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6, e4958. 10.7717/peerj.4958. [PubMed: 29888139]
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. 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. 10.1098/rspb.1996.0184. [PubMed: 8858876]
119. Stanyon R, and Galleni L (1991). A rapid fibroblast culture technique for high resolution karyotypes. Bolletino di zoologia 58, 81–83. 10.1080/11250009109355732.
120. Formenti G, Rhie A, Balacco J, Haase B, Mountcastle J, Fedrigo O, Brown S, Capodiferro MR, Al-Ajli FO, Ambrosini R, et al. (2021). Complete vertebrate mitogenomes reveal widespread gene duplications and repeats. Genome Biol 22, 120. 10.1186/s13059-021-02336-9. [PubMed: 33910595]
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. 10.1093/gigascience/giaa153. [PubMed: 33420778]

122. Seppey M, Manni M, and Zdobnov EM (2019). BUSCO: assessing genome assembly and annotation completeness. Methods Mol. Biol 1962, 227–245. 10.1007/978-1-4939-9173-0_14. [PubMed: 31020564]

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

124. Zar JH (1972). Significance testing of the spearman rank correlation coefficient. J. Am. Stat. Assoc 67, 578–580. 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. 10.1038/nature03154.

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

127. Bao W, Kojima KK, and Kohany O (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11. 10.1186/s13100-015-0041-9. [PubMed: 26045719]

128. Mann HB, and Whitney DR (1947). On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist 18, 50–60. 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 MF, Parker BJ, 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. 10.1038/nature10530. [PubMed: 21993624]

132. Chang CC (2020). Data management and summary statistics with PLINK. Methods Mol. Biol 2090, 49–65. 10.1007/978-1-0716-0199-0_3. [PubMed: 31975163]

133. Ardlie KG, Kruglyak L, and Seielstad M (2002). Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet 3, 299–309. 10.1038/nrg777. [PubMed: 11967554]

134. Bejarano D, Martínez R, Manrique C, Parra LM, Rocha JF, 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. 10.1590/1678-4685-GMB-2016-0310. [PubMed: 30088613]

135. Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC, et al. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. Sci. Data 7, 399. 10.1038/s41597-020-00743-4. [PubMed: 33203859]

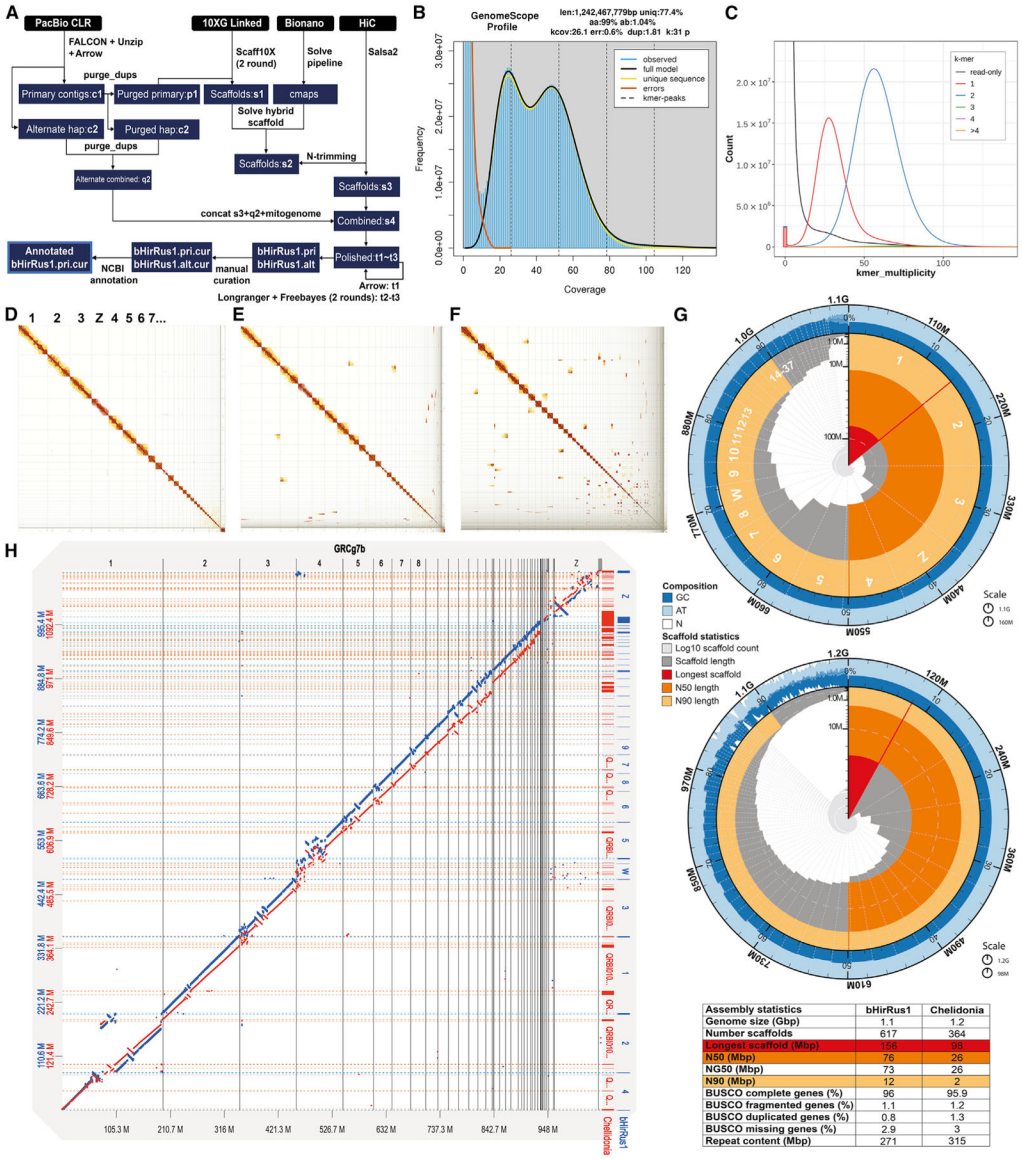
136. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, and Koren S (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res 30, 1291–1305. 10.1101/gr.263566.120. [PubMed: 32801147]

137. Mc Cartney AM, Shafin K, Alonge M, Bzikadze AV, Formenti G, Fungtammasan A, Howe K, Jain C, Koren S, Logsdon GA, et al. (2022). Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nat. Methods 19, 687–695. 10.1038/s41592-022-01440-3. [PubMed: 35361931]

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 10.1080/14786440009463897.

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



Assembly statistics	bHirRus1	Chelidonia
Genome size (Gbp)	1.1	1.2
Number scaffolds	617	364
Longest scaffold (Mbp)	156	98
N50 (Mbp)	76	28
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× (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.

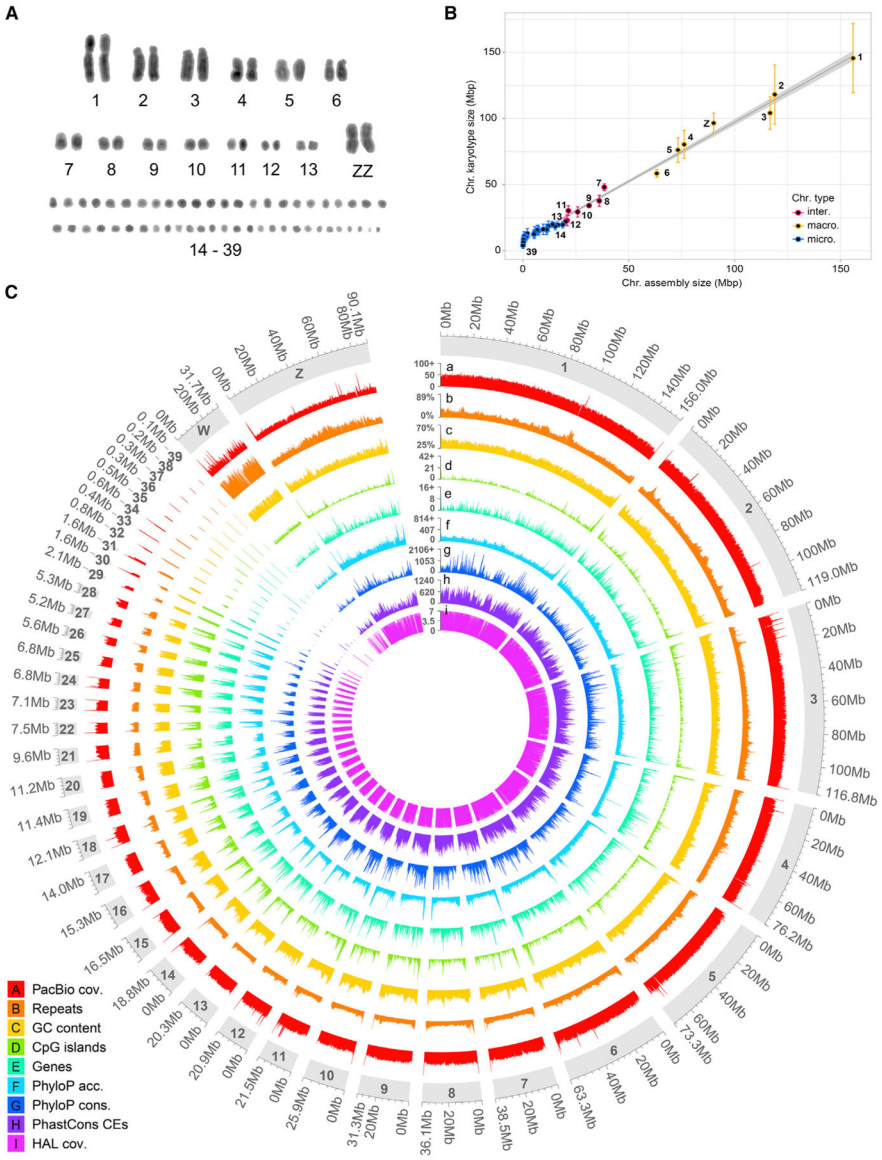


Figure 2. Karyotype reconstruction and reference genome chromosome characteristics

(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.

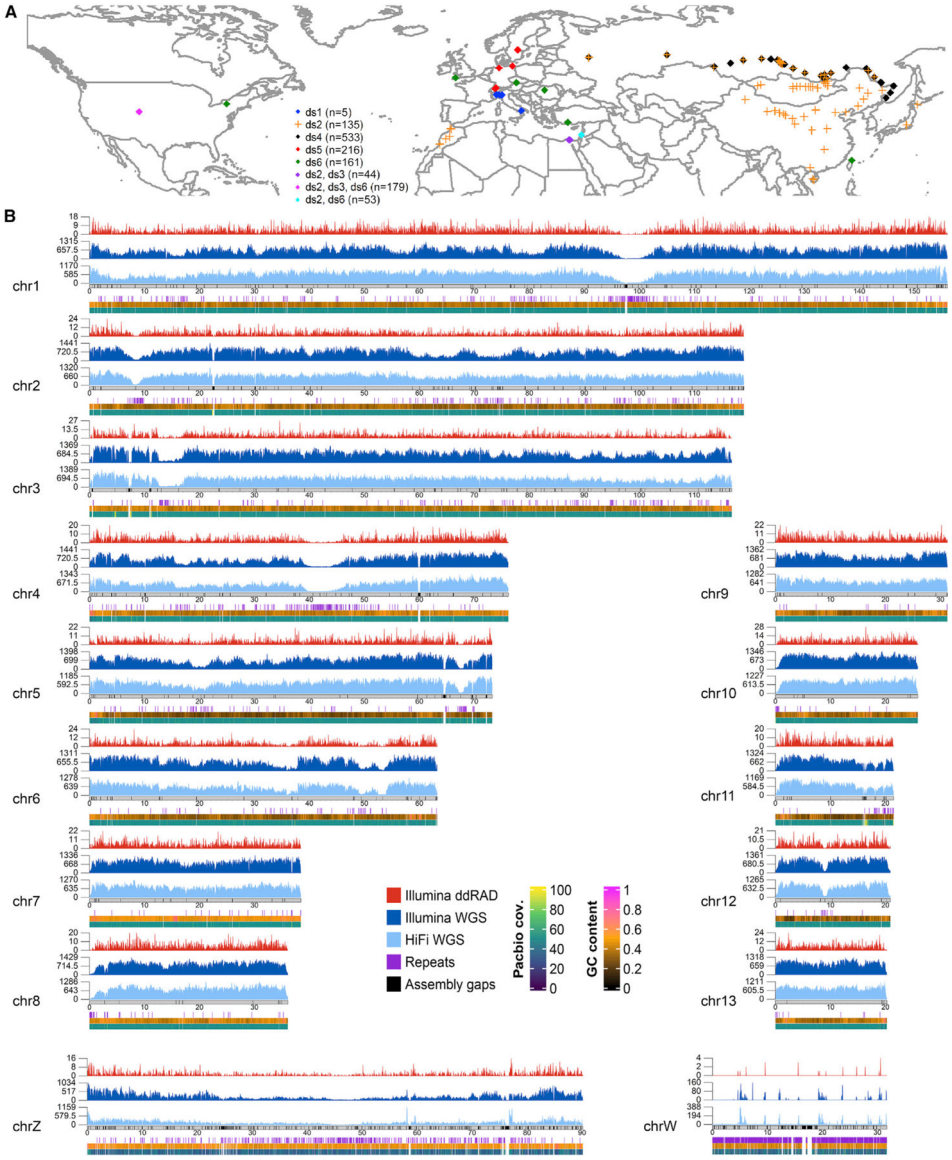


Figure 3. Sampling locations and SNP density per chromosome

(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.

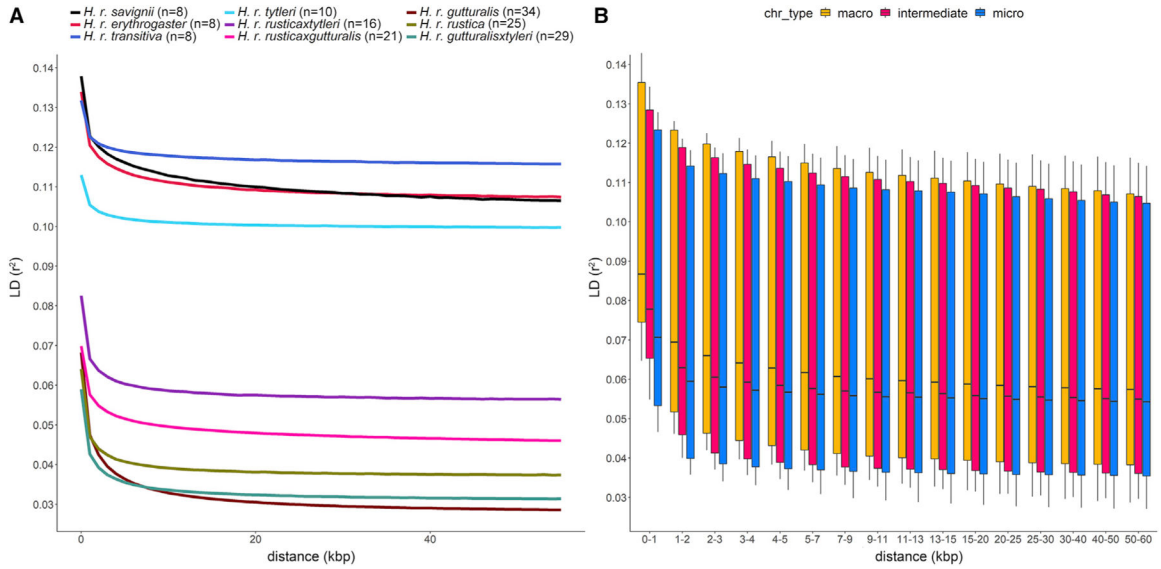


Figure 4. Linkage disequilibrium decay in the barn swallow genome

(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.

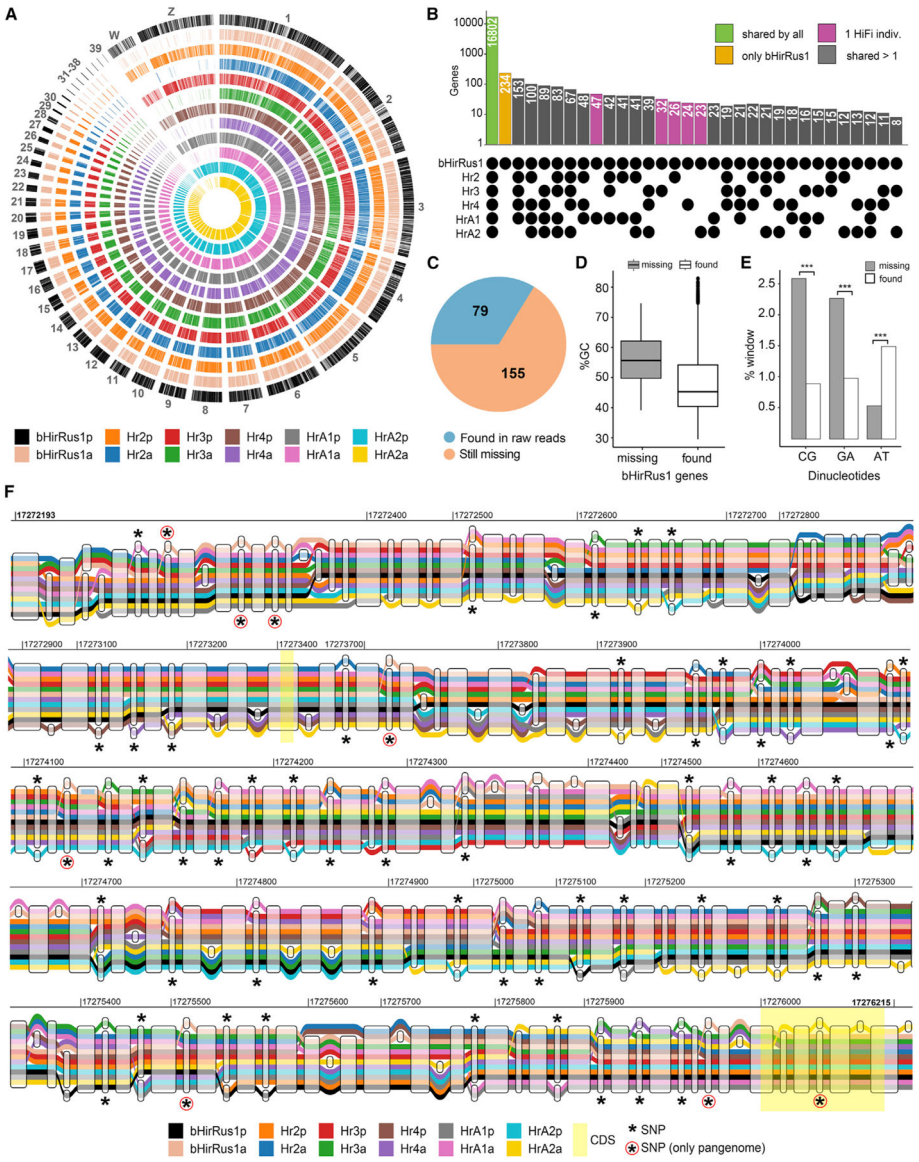


Figure 5. The first pangenome for the barn swallow

(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.

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 Ronn 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.

REAGENT or RESOURCE	SOURCE	IDENTIFIER
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
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/

REAGENT or RESOURCE	SOURCE	IDENTIFIER
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
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)	Thorvaldsdottir 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 etal.[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

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 etal.[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 etal.[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	Guetal.[111]	https://github.com/jokergoo/circlize
ComplexHeatmap	Guetal.[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