Pangenomics provides insights into the role of synanthropy in barn swallow evolution

Author list

Simona Secomandi* (simona.secomandi@unimi.it), Department of Biosciences, University of Milan, Milan, Italy

Guido Roberto Gallo* (guido.gallo@unimi.it), Department of Biosciences, University of Milan, Milan, Italy

Marcella Sozzoni (marcella.sozzoni@studenti.unimi.it), Department of Biosciences, University of Milan, Milan, Italy

Alessio Iannucci (alessio.iannucci@unifi.it), Department of Biology, University of Florence, Sesto Fiorentino (FI), Italy

Elena Galati (elena.galati@unimi.it), Department of Biosciences, University of Milan, Milan, Italy

Linelle Abueg (labueg@rockefeller.edu), Vertebrate Genome Lab, The Rockefeller University, New York City, USA

Jennifer Balacco (jbalacco@rockefeller.edu), Vertebrate Genome Lab, The Rockefeller University, New York City, USA

Manuela Caprioli (manuela.caprioli@unimi.it), Department of Environmental Sciences and Policy, University of Milan, Milan, Italy

William Chow (wc2@sanger.ac.uk), Wellcome Sanger Institute, Cambridge, UK

Claudio Ciofi (claudio.ciofi@unifi.it), Department of Biology, University of Florence, Sesto Fiorentino (FI), Italy

Joanna Collins (jcc@sanger.ac.uk), Wellcome Sanger Institute, Cambridge, UK

Olivier Fedrigo (ofedrigo@rockefeller.edu), Vertebrate Genome Lab, The Rockefeller University, New York City, USA

Luca Ferretti (luca.ferretti@unipv.it), Department of Biology and Biotechnology "L. Spallanzani", University of Pavia, Pavia, Italy

Arkarachai Fungtammasan (chai@dnanexus.com), DNAnexus Inc, USA

Bettina Haase (bet.ha@gmx.de), Vertebrate Genome Lab, The Rockefeller University, New York City, USA

Kerstin Howe (kj2@sanger.ac.uk), Wellcome Sanger Institute, Cambridge, UK

Woori Kwak (woori@hoonygen.com), Hoonygen, Seoul, Republic of Korea; Hoonygen, Seoul, Korea

Gianluca Lombardo (gianluca.lombardo01@universitadipavia.it), Department of Biology and Biotechnology "L. Spallanzani", University of Pavia, Pavia, Italy

Patrick Masterson (patrick.masterson@nih.gov), National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Graziella Messina (graziella.messina@unimi.it), Department of Biosciences, University of Milan, Milan, Italy

Anders Pape Møller (anders.moller@universite-paris-saclay.fr), Ecologie Systématique Evolution, Université Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, Orsay Cedex, France

Jacquelyn Mountcastle (jmountcast@rockefeller.edu), Vertebrate Genome Lab, The Rockefeller University, New York City, USA

Timothy A. Mousseau (mousseau@sc.edu), Department of Biological Sciences, University of South Carolina, Columbia, SC, 29208, USA

Joan Ferrer-Obiol (joan.ferrer@unimi.it), Department of Environmental Sciences and Policy, University of Milan, Milan, Italy

Anna Olivieri (anna.olivieri@unipv.it), Department of Biology and Biotechnology "L. Spallanzani", University of Pavia, Pavia, Italy

Arang Rhie (arang.rhie@nih.gov), Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome, National Human Genome Research Institute, National Institutes of Health (Bethesda, Maryland, USA)

Diego Rubolini (diego.rubolini@unimi.it), Department of Environmental Sciences and Policy, University of Milan, Milan, Italy

Marielle Saclier (marielle.saclier@pasteur.fr), Department of Developmental and Stem Cell Biology, Institut Pasteur/CNRS UMR3738, Cellular Plasticity and Disease Modelling, 25 Rue du Docteur Roux, 75015 Paris

Roscoe Stanyon (roscoe.stanyon@unifi.it), Department of Biology, University of Florence, Sesto Fiorentino (FI), Italy

David Stucki (dstucki@pacificbiosciences.com), Pacific Biosciences, Menlo Park, CA, USA

Françoise Thibaud-Nissen (thibaudf@ncbi.nlm.nih.gov), National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

James Torrance (james.torrance@sanger.ac.uk), Wellcome Sanger Institute, Cambridge, UK

Antonio Torroni (antonio.torroni@unipv.it), Department of Biology and Biotechnology "L. Spallanzani", University of Pavia, Pavia, Italy

Kristina Weber (kweber@pacb.com), Pacific Biosciences, Menlo Park, CA, USA

Roberto Ambrosini (roberto.ambrosini@unimi.it), Department of Environmental Sciences and Policy, University of Milan, Milan, Italy.

Andrea Bonisoli-Alquati (aalquati@cpp.edu), Department of Biological Sciences, California State Polytechnic University - Pomona, Pomona, CA, USA

Erich D. Jarvis (ejarvis@rockefeller.edu), The Rockefeller University (New York, NY, USA), Vertebrate Genome Laboratory, and HHMI

Luca Gianfranceschi† (luca.gianfranceschi@unimi.it), Department of Biosciences, University of Milan, Milan, Italy

Giulio Formenti† (gformenti@rockefeller.edu), The Rockefeller University (New York, NY, USA), Vertebrate Genome Laboratory, and HHMI

* Co-first authors.
† Co-corresponding authors.

Abstract

Insights into the evolution of non-model organisms are often limited by the lack of reference genomes. As part of the Vertebrate Genomes Project, we present a new reference genome and a pangenome produced with High-Fidelity long reads for the barn swallow *Hirundo rustica*. We then generated a reference-free multialignment with other bird genomes to identify genes under selection. Conservation analyses pointed at genes enriched for transcriptional regulation and neurodevelopment. The most conserved gene is *CAMK2N2*, with a potential role in fear memory formation. In addition, using all publicly available data, we generated a comprehensive catalogue of genetic markers. Genome-wide linkage disequilibrium scans identified potential selection signatures at multiple loci. The top candidate region comprises several genes and includes *BDNF*, a gene involved in stress response, fear memory formation, and tameness. We propose that the strict association with humans in this species is linked with the evolution of pathways typically under selection in domesticated taxa.

Keywords

Genome assembly, comparative genomics, pangenomics, genetic markers, positive selection, synanthropy

Introduction

The association with anthropogenic environments includes different degrees of dependence, starting with synanthropy, when a species continues to live in areas occupied and altered by humans (1,2), and ending with domestication, when humans directly control selective pressures. Domestication has been extensively studied in birds (3) and mammals (4,5), where it has been linked to modifications of behavioural mechanisms, particularly a reduction in fear and reactive aggression responses and increased tameness, presumably related to alterations of specific physiological and developmental processes. Among these, neural crest cells development (6,7), corticosteroid hormones release (8) and other stress tolerance-related pathways, such as the glutamatergic signaling (9), are well-documented. Synanthropic species have adapted to exploit human environments without the need of an obligate dependency on anthropogenic resources (10). Typical adaptations are related to immune system response (11), resistance to pollutants (12–14), dietary (15), and behavioural changes (16). Because of its strong association with humans, the barn swallow (*Hirundo rustica*) is a well-suited model to investigate the evolution and genetic bases of behaviours correlated to synanthropy. The barn swallow is a well-studied (17–22) migratory passerine bird with six recognised subspecies in Europe, Asia, Africa and the Americas (23). While still poorly understood, recent studies have started to shed light on its genetics (18,24–29). The barn swallow demographic history reconstructed from genomic data suggests that the current barn swallow

distribution derives from a relatively recent expansion, probably driven by the spread of human settlements providing more nesting opportunities (28,30). In the European subspecies, no evidence of population structure was observed, likely due to extensive gene flow between breeding populations (25). Studies on the barn swallow genomic architecture and adaptations have been limited by the lack of a highly contiguous, complete, and well-annotated reference genome for the species. The first reference genome, released in 2019 by our research group (31), was a scaffold-level assembly for the Eurasian subspecies generated combining PacBio long-read sequencing (32) and Bionano Direct Label and Stain (DLS) optical mapping (33). The second was a fragmented assembly for the same subspecies based on Illumina short reads and released in 2020 by the B10K Consortium (34,35). Here we present the first chromosome-level assembly for the Eurasian barn swallow *H. r. rustica*, generated using the Vertebrate Genomes Project (VGP) assembly pipeline (36), and the first pangenome for the species to expand the characterization of its intraspecific variation (37,38). With this assembly we identified conserved and accelerated genomic regions in the barn swallow genome, and generated a catalogue of genetic markers to detect high-linkage disequilibrium (LD) regions. Both approaches pointed at candidate genes known to be implicated in stress response, fear memory formation and vocal learning in songbirds (39). These processes are associated with tameness and domestication in birds (9), suggesting that the synanthropic habits of barn swallows could have evolved through similar selective pressures and pathways as those shaping the evolution of domestic taxa.

Results and discussion

A reference genome for the barn swallow. Using the VGP genome assembly pipeline v1.6 (36) (Additional file 1: Figure S1), we generated the first chromosome-level reference genome assembly ('bHirRus1' hereafter) and an alternative-haplotype assembly for the barn swallow. This included generating contigs with PacBio CLR long reads and scaffolding them with 10x linked reads, Bionano optical maps, and Hi-C reads. We also generated a mitochondrial genome for the species (Additional file 1: Figure S2, Supplementary Note). We sequenced a female, to obtain both Z and W sex chromosomes. After our manual curation (Supplementary Note) the primary assembly was 1.11 Giga base pairs (Gbp) long, with a scaffold NG50 of 73 Mega base pairs (Mbp) and a per-base consensus accuracy of Q43.7 (~0.42 base errors/10 kilo base pairs, kbp; Additional file 1: Tables S1-2, see Supplementary Note for the extended evaluation). We assigned 98.2% of the assembled sequence to 39 autosomes and the Z and W sex chromosomes (Additional file 1: Figure S3a, Table S3), which are usually challenging to assemble due to their highly repetitive nature (40). The assembly exceeds the VGP standard metrics (6.7.Q40.C90) (36). The chromosome reconstruction (2n = 80) matches our cytogenetic analysis (Fig. 1a; Additional file 1: Supplementary Note), and is in line with the current

literature on pachytene karyotypes of the barn swallow (41). Based on the original chicken chromosome classification (42) and our chromosome sizes (Additional file 1: Table S3, Supplementary Note), we define chromosomes 1-6 and Z as macrochromosomes, 7-13 and W as intermediate-size chromosomes, and 14-39 as microchromosomes. The size of the assembled chromosomes tightly correlates with the size of the chromosomes estimated from karyotype images (Spearman's $\rho = 0.99$, $n = 40$, $P < 2.2 \times 10^{-16}$, Fig. 1b, Additional file 1: Table S4). As expected (36), PacBio long-reads coverage shows haploid coverage for Z and W (Fig. 1c track A). The total repeat content of the assembly is 271 Mbp (22.9%, Fig. 1c track B, Additional file 1: Table S3), in line with Genomescope2.0 (43) predictions (Additional file 1: Figure S4a, Table S5). The GC content is 42.5% (Fig. 1c track C, Additional file 1: Table S3). Functional gene completeness, measured with BUSCO (44), is 96% (Additional file 1: Figure S5a, Table S6).

Functional annotation. Newly-generated IsoSeq and RNAseq data, RNAseq data from other individuals (45) (Table S7), 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). The NCBI Eukaryotic genome annotation pipeline (36,46) identified 18,578 genes and pseudogenes, of which 15,516 were protein coding. Among these, 15,130 (97.5%) aligned to UniProtKB/Swiss-Prot curated proteins, covering $\geq$ 50% of the query sequence, while 10,797 (69.6%) coding sequences aligned for $\geq$ 95%. In line with other birds (47), ~52% of the total bp is annotated as genes, of which ~90% are annotated as introns and ~5% as coding DNA sequences (Additional file 1: Table S8).

Chromosome size and genomic content. Differences in GC, CpG islands, gene and repeat content between bird macro-, intermediates and microchromosomes are likely the product of the evolutionary process that led to stable chromosome types in birds (48). Similarly to the zebra finch *Taeniopygia guttata (49)* genome, bHirRus1 chromosome size negatively correlates with GC content (Spearman's $\rho = -0.972$, $n = 40$, $P < 2.2 \times 10^{-16}$), CpG island density (Spearman's $\rho = -0.925$, $n = 40$, $P < 2.2 \times 10^{-16}$), gene density (Spearman's $\rho = -0.364$, $P < 2.5 \times 10^{-2}$) and repeat density (Spearman's $\rho = -0.51$, $n = 40$, $P = 1.2 \times 10^{-3}$; Fig 1c; Additional file 1: Figure S6). Indeed, microchromosomes are GC-rich (Wilcoxon test, $W = 0$, $P = 1.4 \times 10^{-7}$), CpG-rich (Wilcoxon test, $W = 3$, $P = 2.2 \times 10^{-7}$), gene-rich (Wilcoxon test, $W = 94$, $P = 9.9 \times 10^{-3}$) and repeat-rich (Wilcoxon test, $W = 103$, $P = 2.0 \times 10^{-2}$) with respect to the other types of chromosomes.

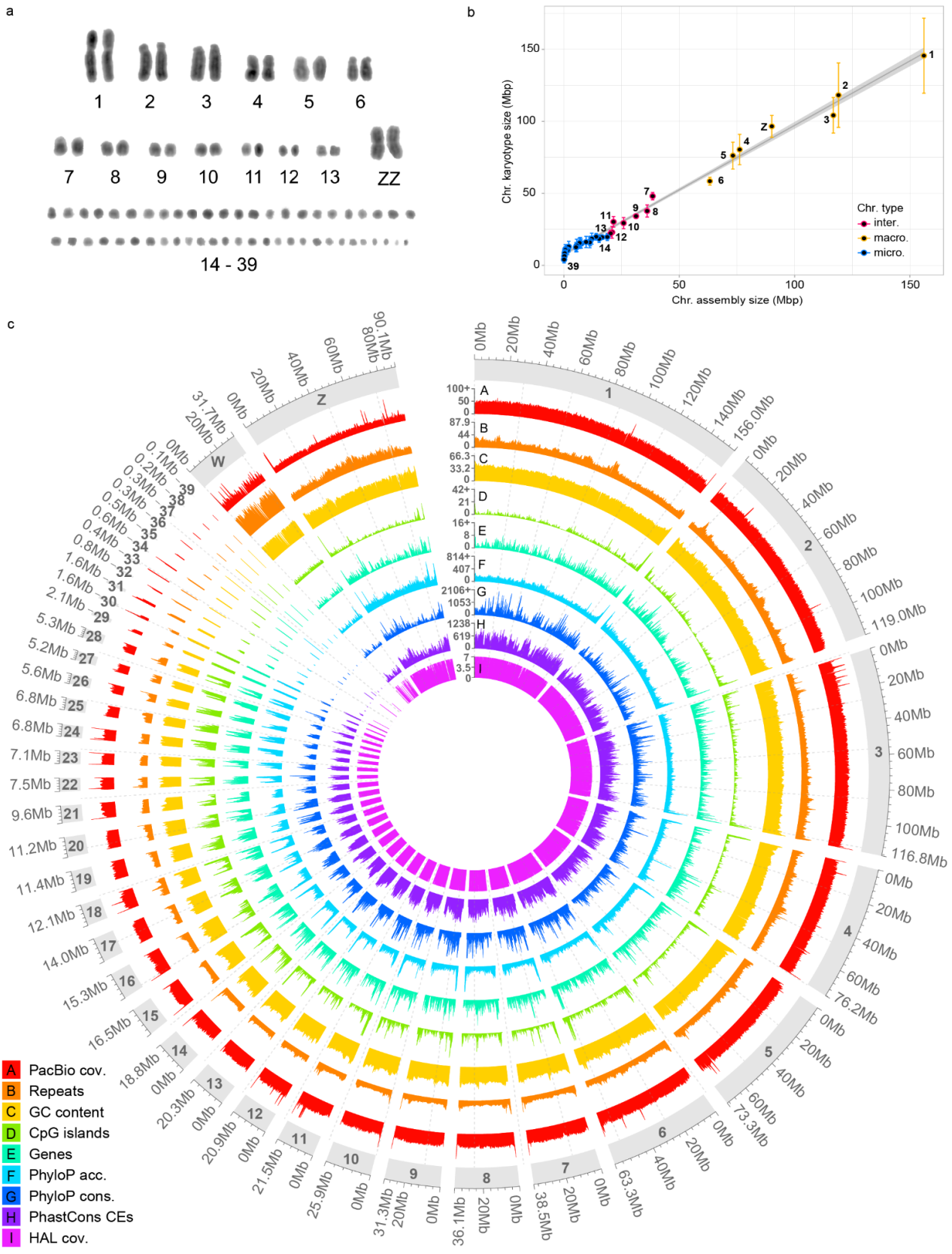

Fig. 1 Karyotype reconstruction and chromosome-level assembly for the barn swallow. **a** DAPI-stained karyotype of a male Eurasian barn swallow (inverted colours). **b** Correlation between assembled chromosome length (x) and the estimated chromosome length from karyotype images (y). W sex chromosome 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 for visualisation whenever

necessary (marked with +). PacBio long-read coverage (A); % repeat density (B); % GC (C); CpG islands density (D); gene density from NCBI annotation (E); accelerated sites density computed with phyloP (F); conserved sites density computed with phyloP (G); conserved elements (CEs) density computed from phastCons analysis (H); coverage of bHirRus1 in the Cactus HAL alignment, i.e. number of species aligned (I).

Comparison between bHirRus1 and the previous scaffold-level assembly. Compared to the previous assembly (here after 'Chelidonia', scaffold NG50 26 Mbp; Additional file 1: Table S1), the VGP assembly pipeline and our subsequent manual curation increased the assembly contiguity to the chromosome level (see Supplementary Note for an extended comparison). Assembly QV is also considerably increased (43.7 vs 34; Additional file 1: Table S2). The repeat content decreased from 315 Mb to 271 Mb (Additional file 1: Figure S5c). BUSCO completeness increased in bHirRus1 (96% vs 95.9%), and BUSCO genes were less duplicated (0.8% vs 1.3%) and less fragmented (1.1% vs 1.2%; Additional file 1: Figure S5c, Table S6). We reconciled the larger size of Chelidonia (1.2 Gbp; Table S1) with the size of bHirRus1 (1.11 Gbp) by identifying 55 Mbp of repeats, sequence overlaps, low-coverage regions and haplotigs in Chelidonia (Additional file 1: Table S9, Supplementary Note).

Reference-free whole-genome multiple species alignment and selection analysis. To identify regions under positive and negative (purifying) selection, we generated a reference-free, whole-genome multiple alignment using Cactus (50,51). The alignment included bHirRus1, six publicly-available chromosome-level Passeriformes genomes, and the chicken GRCg7b genome (Additional file 1: Figure S7, Table S10). Most of the species are synanthropic, domesticated or live partially in contact with humans. Overall, the coverage of the alignment with bHirRus1 was uniform in macrochromosomes, intermediate chromosomes, with the exception of chromosome W, and the largest microchromosomes (Fig. 1 track I; Additional file 1: Table S11). The mean alignability between all the species and the barn swallow was ~76% (Additional file 1: Table S10). Using a 4-fold-degenerate sites neutral model and the Cactus alignment, we found that 0.96% of bHirRus1 bases are accelerated (i.e, evolve at higher rate than that under neutral evolution) and 2.71% are conserved (i.e, evolve at a lower rate) using phyloP with false-discovery rate (FDR) correction (52) (Fig. 1c track F-G; Additional file 1: Figure S8, Table S12). Approximately 52% and 63% of accelerated and conserved nucleotides, respectively, fell within genes (Additional file 1: Figure S8e, Table S12). Only ~0.9% and ~17% of accelerated and conserved bases overlapped with coding sequences (CDS), in line with previous studies (53,54). Using phastCons (55) and an *ad hoc* parameters set (coverage and smoothing), we identified ~3 million conserved elements (CEs) covering 12.3% of the barn swallow genome (133 Mbp; Fig. 1c track H; Additional file 1: Table S12). Similarly to the phyloP analysis, significant overlaps were observed between CEs and genes (~61%), with ~14% of CEs overlapping CDS (Additional file 1: Figure S8e, Table S12), as expected (53,54). While conserved sites density was

weakly positively correlated with chromosome sizes (Spearman's $\rho = 0.35$, $n = 40$, $P < 3.4 \times 10^{-2}$), without significant differences between chromosome classes (Wilcoxon test, $W = 244$, $P = 0.189$), accelerated sites density was strongly negatively correlated with chromosome size (Spearman's $\rho = -0.80$, $n = 40$, $P < 9.5 \times 10^{-8}$), with microchromosomes richer in accelerated sites than the other chromosomes (Wilcoxon test, $W = 50$, $P = 4.6 \times 10^{-5}$), as already observed in other birds (56). The Gene Ontology (GO) analysis on the top 5% genes with highest overlapping with phyloP accelerated sites (Additional file 1: Table S13) did not disclose any enriched GO term (Additional file 1: Table S14, Supplementary Note). PhyloP conserved sites showed a highly significant positive correlation with CEs detected with phastCons (Spearman's $\rho = 0.83$, $n = 108010$, $P < 2.2 \times 10^{-16}$; Fig. 1c). Since phyloP sites can be considered a higher confidence subset within the larger phastCons set, we focussed our subsequent analyses on phyloP sites. As expected, CDS were the most conserved (57) (Additional file 1: Figure S8c, Table S12). The GO analysis on the top 5% genes with highest overlapping between CDS and phyloP conserved bases (Additional file 1: Table S15) revealed an enrichment for genes involved in DNA-binding, transcriptional regulation and nervous system development (Additional file 1: Table S16). The top 20 genes were largely involved in neural development and differentiation (Additional file 1: Table S15, Supplementary Notes). The top candidate was *CAMK2N2* (89% CDS bases conserved; Additional file 1: Table S15), located on chromosome 10. In the Cactus alignment, in correspondence with its CDS coordinates, all the species have the same base composition, with the exception of the chicken, which has a few SNPs (Additional file 1: Figure S9). PhyloP conserved bases were located only in regions without SNPs, while phastCons CEs comprise also regions which are not fully conserved between all species. *CAMK2N2* encodes a protein that acts as an inhibitor of calcium/calmodulin-dependent protein kinase II (*CAMKII*). *CAMKII* has a vital role in long-term potentiation of synaptic strength (LTP) and learning, via regulation of glutamate receptors (AMPA) (58–62). *CAMKII* is also one of the main calcium/calmodulin targets after the activation of NMDA (N-methyl-d-aspartate) glutamate receptors, which are involved in memory formation (63). Moreover, a peptide derived from *CAMK2N2 (*tatCN21) impairs fear memory formation by blocking *CAMKII* activity (64), and overexpression of *CAMK2N2* in the hippocampus was found involved in memory formation (65). In the Bengalese finch *Lonchura striata domestica (9)*, one of the species included in the Cactus alignment, the glutamatergic system contributed to the attenuation of stress response and aggressive behaviour under domestication. Finally, in high stress lines of the domesticated Japanese quail *Coturnix japonica*, *CAMK2N2* and *CAMKII* have been detected as deleted, together with other genes in the same networks, compared with low stress lines (66,67). Loss of genes in this network may be responsible for the reduced growth rate and low basal weight of the high stress quails (67). Since *CAMK2N2* is likely involved in behavioural and physiological changes under domestication in

birds, we evaluated it in relation to the onset of synanthropic habits in the barn swallow. We generated an alignment of transcripts from 38 species (17 domesticated or synanthropic, 21 wild; Additional file 1: Table S17). However, we did not observe any pattern specific to domesticated or synanthropic species, and the single-gene phylogenetic tree substantially matched the known phylogeny. Thus, any role of *CAMK2N2* in synanthropic habits or domestication would have to be ascribed to non-coding regulatory elements. In vocal learning bird species, domestication was also found involved in the control of dopaminergic signalling in neural circuits that are crucial for vocal learning (9). Among the top 20 genes with the most overlap between CDS and phyloP conserved bases (Additional file 1: Table S15), *FOXP2* has 74% of its CDS bases conserved. This gene received great attention for its role in language and speech, since mutations in its sequence cause, among others, speech impairments (68–73). In the zebra finch, a vocal learner like the barn swallow, this gene has a marked expression in brain regions involved in song learning (74–77). Another candidate gene detected and previously associated with song learning is *UBE2D3* (75% CDS conserved; Additional file 1: Table S15), a gene located in a region of the human genome associated with musical abilities (78–80), which include recognizing, reproducing and memorising sounds. *CAMK2N2*, *FOXP2* and *UBE2D3* were also in the top 5% genes with the most overlaps between CDS and CEs bases detected with phastCons (Additional file 1: Table S18).

Towards 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 (81). To reduce bias towards a single reference genome, we generated high-coverage (~15-30x) HiFi whole-genome sequencing (WGS) data for five additional *H. r. rustica* individuals (Additional file 1: Tables S19-S20), assembled them with Hifiasm (82), and used both primary and alternate haplotypes (Additional file 1: Table S21) to generate the first pangenome variation graph (37,38) for the species (Fig. 2, Additional file 1: Figure S10). The HiFi-based primary assemblies had a contig NG50 between 2-8.6 Mbp, while the alternate between 0.2-1.7 Mbp, proportional to sequencing coverage (Additional file 1: Tables S20; S21). All primary assemblies shared 90.5% of their sequence (core genome), while all the HiFi individuals, considering both primary and alternate, shared 92.6% of bHirRus1 genes (Fig. 2c-d; Additional file 1: Table S22). 1.36% (236) of bHirRus1 annotated genes were not found in the HiFi assemblies (Fig. 2d; Additional file 1: Tables S22-23). Of those genes, 79 were found in the raw-reads of at least one individual for > 80% of their sequence with > 99% identity (Additional file 1: Table S23). The absence of the remaining 157 genes (0.87%) from both HiFi raw reads and HiFi-based assemblies, may either be due to the known GA dropout in HiFi reads (83), or to real gene losses in those individuals.

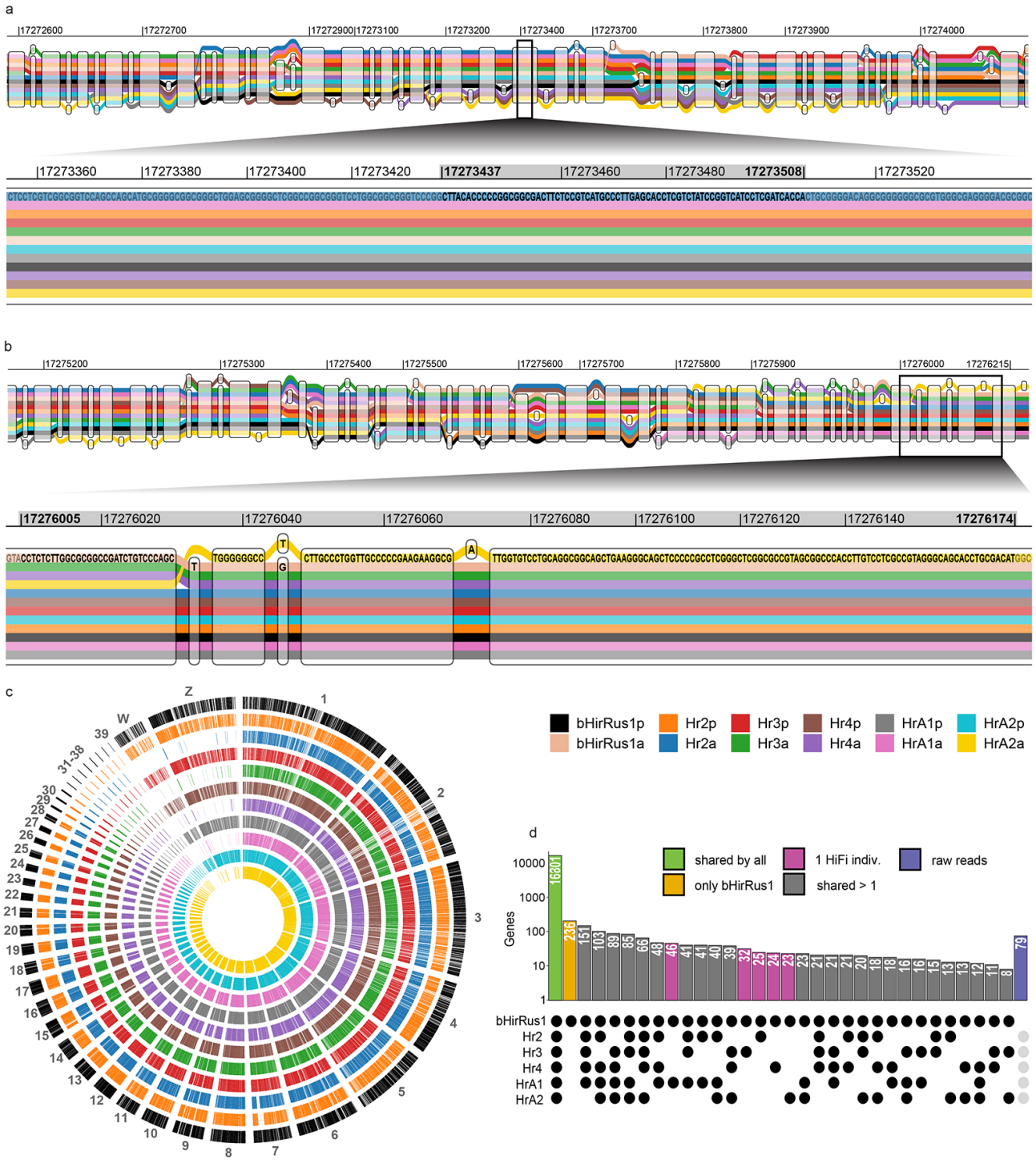

Fig. 2 The first pangenome for the barn swallow. **a** *CAMK2N2* initial region in the barn swallow pangenome. bHirRus1 Chr10 ('bHirRus1p') is shown together with the alternate assembly 'bHirRus1a', the five HiFi-based primary (Hr2p, Hr3p, Hr4p, HrA1p, HrA2p), and their alternate assemblies (Hr2a, Hr3a, Hr4a, HrA1a, HrA2a). The zoomed part shows the first CDS (grey rectangle, 17,273,437-17,273,508). **b** *CAMK2N2* terminal region. The zoomed part shows the details of the second CDS (grey rectangle, 17,276,005-17,276,174). **c** Circos plot showing the annotated genes of bHirRus1 (primary, black) and orthologs in the other individuals (primary and alternate combined). Tracks follow the same colour legend as a and b. **d** Presence or absence of bHirRus1 genes in the other individuals included in the pangenome. The histogram reports the number of genes shared between bHirRus1 (primary) and each of the other individuals or groups of individuals (primary and alternate assemblies combined). The majority of the genes are shared between all individuals (green), while only 236 genes are exclusive of bHirRus1 (yellow). Genes shared only between bHirRus1 and another individual are shown in purple. The remaining bHirRus1 genes were found in 2 or more individuals (grey). Seventy-nine out of the 236 genes exclusive of bHirRus1 were found with BLAST (84) in at least one individual HiFi reads (violet), and therefore not properly assembled in the HiFi-based assemblies.

Marker catalogue and genome-wide density. In parallel to our phylogenomic analyses, we used bHirRus1 as reference and our high coverage HiFi WGS dataset (ds1, ~20x coverage, N = 5) to generate a comprehensive catalogue of single-nucleotide polymorphisms (SNPs; Additional file 1: Supplementary Note). We complemented this information with all the publicly available genomic data for the species (Additional file 1: Figure S11, Table S24), including two Illumina WGS datasets (28,29) (ds2 and ds3.1, ~6.8x, N = 159) and four ddRAD datasets (24,25,27,28) (ds3.2 through ds6, ~0.07x; N = 1,162). Despite the fewer individuals in HiFi WGS, the average SNP density and distribution (Fig. 3, light blue track; 142.37 SNPs/10 kbp; Additional file 1: Table S25) was comparable to the one computed for Illumina WGS (Fig. 3, dark blue track; 160.34 SNPs/10 kbp; Additional file 1: Table S25), suggesting that this sequencing method yields a high and accurate reads mappability even when only small datasets are available. We also performed a coverage titration experiment (Additional file 1: Supplementary Note) and found that SNP distribution was still uniform across chromosomes even when HiFi WGS were downsampled to 5x (96.33 SNPs/10 kbp; Additional file 1: Figure S12, Table 25). Chromosome W showed the lowest SNP density among all chromosomes (HiFi WGS 3.16 SNPs/10 kbp; HiFi WGS 5x 1.01 SNPs/10 kbp; Illumina WGS 1.38 SNPs/10 kbp), in line with the fact that chromosome W is present as single copy only in females (the heterogametic sex), and it has the highest content of heterochromatin and repeat elements, hindering variant calling (85). In contrast, we identified a higher number of SNP markers on chromosome Z (HiFi WGS 31.8 SNPs/10 kbp; HiFi WGS 5x 2.34 SNPs/10 kbp; Illumina WGS 53.3 SNPs/10 kbp). As expected, ddRAD exhibited very localised peaks of SNP density (0.8 SNPs/10 kbp; Fig. 3, 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 (Additional file 1: Figure S13). Opposed to previous findings in humans (86,87), we detected a positive correlation between chromosome GC content and SNP density in all datasets (Additional file 1: Supplementary Note).

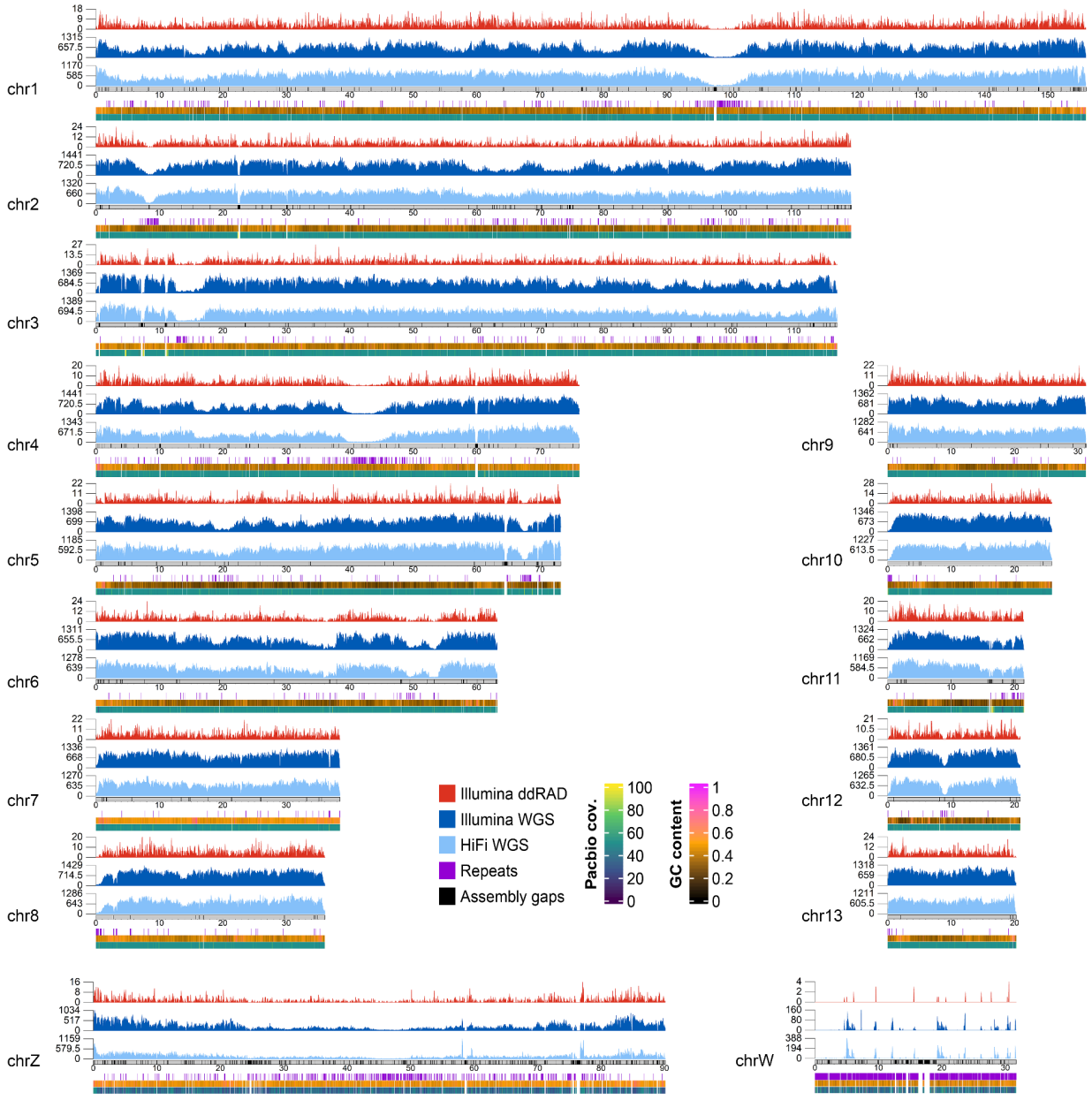

Fig. 3 SNPs density per chromosome. Macrochromosomes and intermediate chromosomes are shown here, while microchromosomes are shown in Extended Data Fig. 7. SNP density, coloured according to the different types of genomic data used, was computed over 40 kbp windows. The numbers on the y axis of each density track indicate the maximum and average values of SNP density for each track. Light blue: HiFi WGS data (ds1). Dark blue: Illumina WGS data from ds2 and ds3.1. Red: Illumina ddRAD data from ds3.2 through ds6.8. All available samples from the same sequencing technology were considered together. Additional tracks in the lower panel show repetitive regions of the genome (violet bars; only regions larger than 3 kbp are plotted), GC content and PacBio reads coverage. Grey ideograms represent chromosomes in scale, with assembly gaps highlighted as black bars.

Genome-wide linkage disequilibrium. LD reflects the evolutionary history of populations as it can be influenced by selective pressures (88–90), recombination rate (91,92), migration (93), genetic drift (94) and population admixture (95,96). Assessing its decay is pivotal to the success of genome-wide association studies (GWAS) (97,98) because it provides an estimate of the number of molecular markers required to detect significant associations between markers and causative loci. Since WGS is usually very useful to describe LD patterns (92), and no previous study estimated

genome-wide LD decay in the barn swallow, we evaluated LD using the SNPs in our catalogue derived from WGS (ds2 and ds3.1). Genome-wide average r^2 varied between *H. rustica* subspecies (Fig. 4a, Additional file 1: Table S26). As expected (99), absolute r^2 decreased with increasing sample size and marker distance (*H. r. savignii*, *H. r. erythrogaster*, *H. r. transitiva*; Fig. 4a, Additional file 1: Table S26). 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 (25,100). Average r^2 at increasing distance varied also across chromosome types, confirming that avian microchromosomes are characterised by higher rates of meiotic recombination, and thus lower LD, than macrochromosomes (Fig. 4b; Additional file 1: Table S27) (48,101,102).

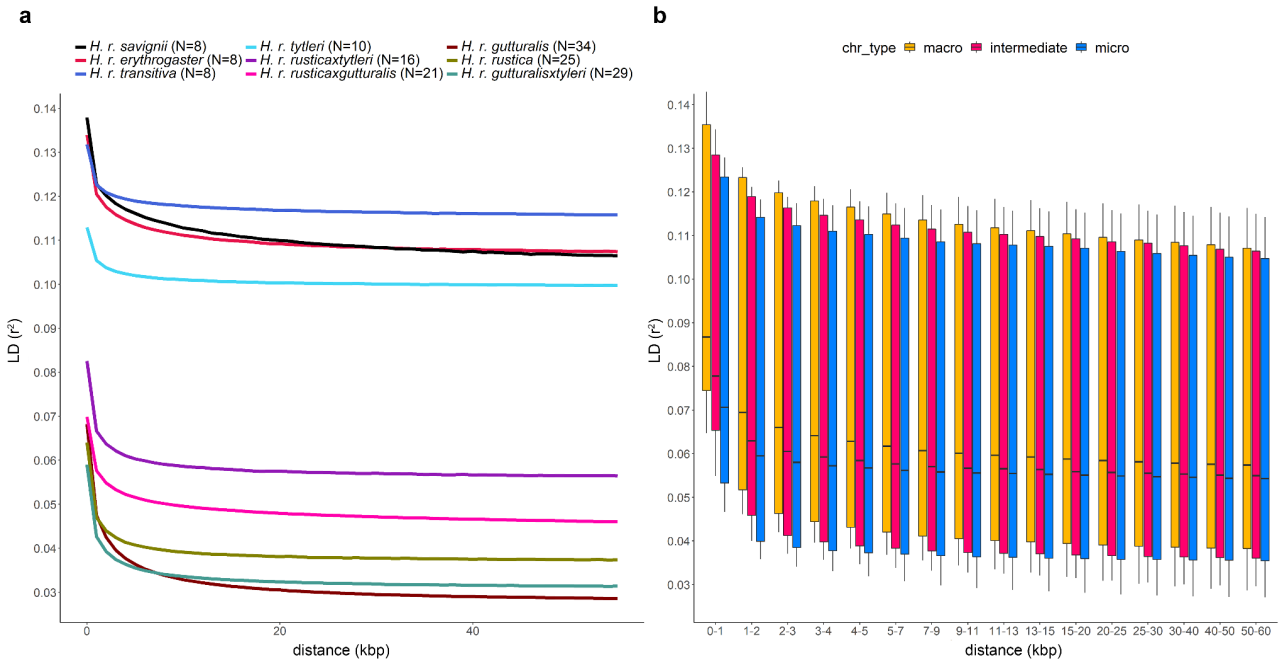

Fig. 4 Linkage disequilibrium decay in the barn swallow genome. **a** Average r^2 values plotted against physical distance (kbp) for the different populations belonging to ds2 and ds3.1 (Illumina WGS data). **b** Average r^2 values in macrochromosomes, intermediate chromosomes, and microchromosomes according to pairwise distance (kbp) between SNPs. LD median estimates were obtained averaging values from all Illumina WGS data populations (ds2 and ds3.1).

Candidate genes in high LD blocks. We performed an initial chromosome scan using Illumina WGS data from the *H. r. erythrogaster* and *H. r. savignii* subspecies (28) (ds3.1) to identify potential regions of interest (ROIs) exhibiting high LD values (average $r^2 > 0.3$). Despite the small sample size and the rapid genome-wide LD decay, our analyses revealed the presence of 78 ROIs, many of which (N = 57/78) spanned at least one annotated protein coding gene (N = 83; Table S28). Excluding ROIs containing sequences potentially collapsed in the reference or not overlapping with annotated genes, the locus showing the highest r^2 values was on chr6 (ROI 45) and harboured four genes (*CCDC34*, *LGR4*, *LIN7C* and *BDNF*; Fig. 5a; Additional file 1: Table S28). Among these genes, *BDNF* is particularly interesting because it

encodes a major neurotrophin involved in neuronal plasticity and differentiation (103,104). In zebra finch males, its transcript is upregulated to high levels in the high vocal centre (HVC) by singing activity (105), particularly when juveniles start to emit vocalisations, and its tissue-specific overexpression significantly increases during sensorimotor song learning (39,106,107). BDNF is also implicated in neural crest cells development (108), and studies in multiple domesticated mammalian species suggest a role for the modification of neural crest development in driving the concerted evolution of tame phenotypes during domestication (i.e., 'domestication syndrome') (6,7). It is also extensively implicated in the response to stress, fear, and fear memory consolidation (109). Similarly to other species (110), barn swallow *BDNF* presents alternative transcripts (Fig. 5b), three of which (transcript variants X2, X3, X4) lead to the same amino acid sequence, suggesting the presence of important regulatory elements. In other bird species, temperature (chicken (111,112)) and prolonged social isolation (zebra finch (112)) affect the expression of *BDNF* through a methylation-mediated mechanism associated with CpG sites located within CpG islands upstream of the translation start site, as well as in the coding region. Initially, using WGS data from American and Egyptian samples (28) (ds3.1), we detected 6 LD blocks comprising 104 SNPs within the *BDNF* gene region. Of these SNPs, 30 directly alter CpG sites, either in the reference or in the alternate allele sequence (Additional file 1: Table S29). The highest LD values were identified within *H. r. savignii* population (Additional file 1: Figure S14a), where we also detected an average homozygosity (i.e. the average proportion of homozygous genotypes) of ~88.8% across all samples for the genotyped SNPs within the gene (Additional file 1: Table S29). The strong LD detected at CpG sites may indicate that certain alleles have been favoured by selection (97,113). In the specific case of the Egyptian barn swallow, where there is evidence of a past bottleneck event (28), we cannot exclude that genetic drift may have also played a role. However, the same genomic region in all other available WGS populations (ds2) had similar LD patterns (Additional file 1: Figure S14). For instance, *H. r. transitiva* showed very high pairwise LD values within *BDNF* gene coordinates (Additional file 1: Figure S14c). We further confirmed the presence of a potential selective signature within this genomic region by computing population haplotype homozygosity statistics (iHS, the integrated haplotype homozygosity score) on chr6 in WGS ds3.1, ds2.1 and ds2.2. The ROI harbouring *BDNF* identified with genome-wide LD scans was associated with significant outlier peaks also in this analysis (Additional file 1: Figures S15-16). Four CpG islands are present within the sequence of *BDNF* in the barn swallow (Fig. 5b, blue blocks). The first CpG island corresponds to one of the two genomic regions containing methylated sites previously described in zebra finch (112). We found that four of the seven CpG sites reported in zebra finch are conserved in the barn swallow (Fig. 5c, highlighted in yellow). One SNP present in our barn swallow markers catalogue (chr6:53,908,036) directly affects a

CpG site adjacent to a zebra finch methylation site (112) (Fig. 5c, SNP adjacent to the first highlighted CpG site). We also analysed this region in the Cactus multialignment and found that all of the zebra finch CpG sites are conserved in all other bird species, except for the chicken, where only two sites are conserved as CpG (Fig. 5c). The presence and conservation of CpG sites in the barn swallow, together with the identified selection signatures associated with this genomic region, reinforce the importance of these sites. CpG islands are known to directly affect the transcription of genes by altering local chromatin structure, mostly through methylation of CpG dinucleotides (111). For *BDNF* methylation-dependent transcriptional regulation involving CpG islands has been shown to affect fear memory consolidation (114), a process strictly involved in domestication. Methylation state assays could potentially help to further investigate the role played by epigenetic modifications of *BDNF* in the barn swallow, providing additional insights on the evolution of tameness-related habits in this species.

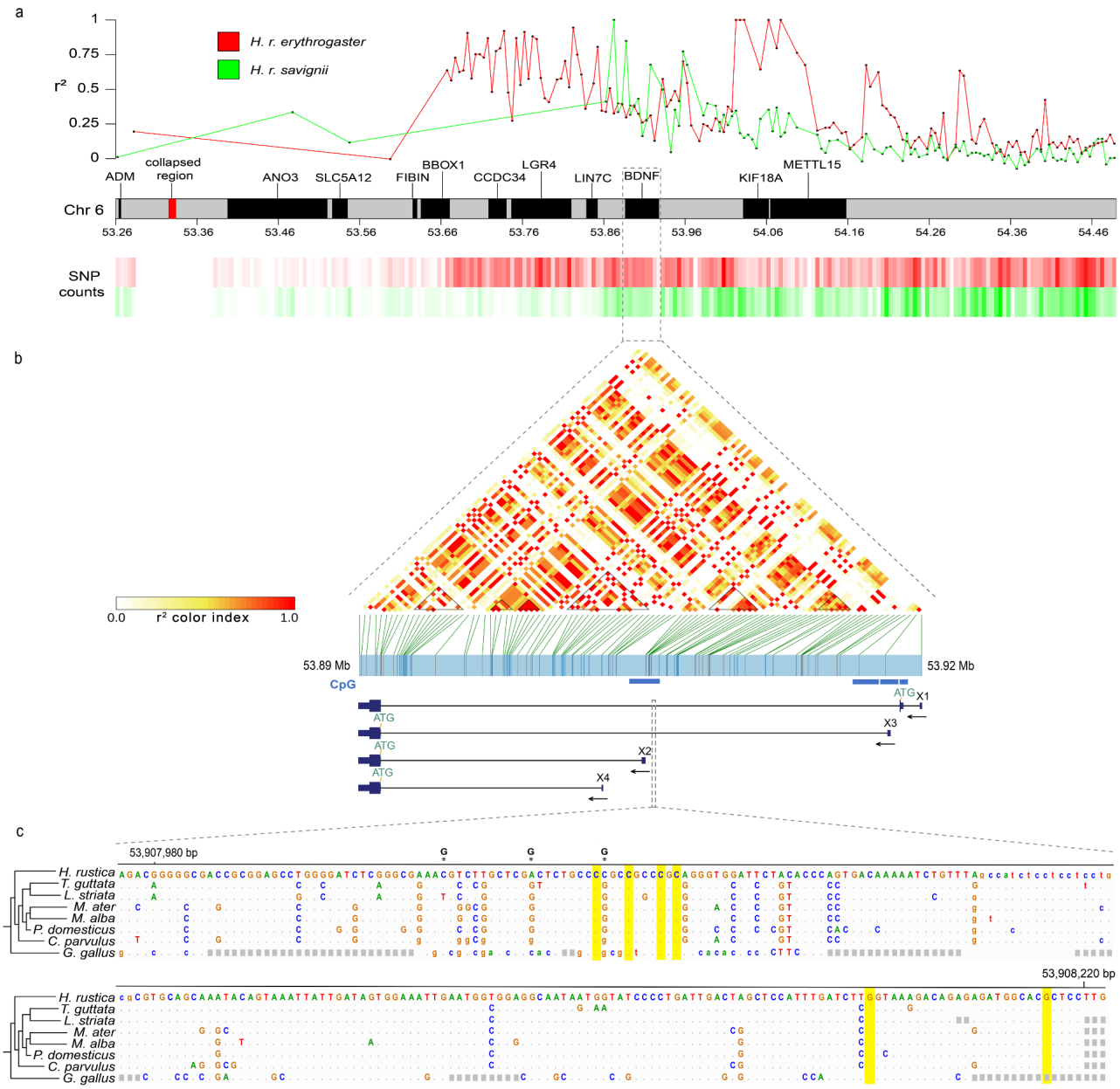

Fig. 5 Patterns of LD blocks in genomic regions on chr6. **a** Average r^2 values computed over 5 kbp windows on chr. 6 (upper panel; from 53.26 Mb to 54.49 Mb) for the *H. r. savignii* (green) and *H. r. erythrogaster* (red) populations (ds3.1). The region shown in the plot extends beyond ROI 45. Each point represents the average r^2 value per window and was placed at the midpoint of the genomic region. The heatmap in the lower panel represents SNP counts for the two populations analysed. **b** Upper panel: LD heatmap within *BDNF* gene coordinates considering the two populations combined. Black triangles indicate LD blocks. Blue horizontal blocks mark the presence of CpG islands. Lower panel: barn swallow *BDNF* four transcript isoforms X1, X2, X3 and X4 (big rectangles: coding exons; small rectangles: non coding exons; horizontal line: introns; arrows indicate the direction of transcription). **c** Cactus multiple alignment of the zebra finch (second line) region containing CpG sites important for methylation-dependent regulation (112). Asterisks: SNPs present in barn swallow marker catalogue. Alternate base is shown on top of the barn swallow reference sequence. Yellow: zebra finch methylated sites (112). The second, third and sixth CpG sites are conserved in the barn swallow. The first one (at position 53,908,035) is not fixed in the barn swallow but the transition of the adjacent polymorphic site from reference (C) to alternate (G) allele leads to the formation of a CpG site.

Conclusion

Using our high-quality, karyotype-validated and fully annotated chromosome-level reference genome for the barn swallow in combination with comparative and population genomics, we detected genes involved in domestication and song learning. Particularly, *CAMK2N2* has a role in fear memory formation and is likely involved in the glutamatergic system, which in turn plays a key role in domestication through the attenuation of stress response and aggressive behaviour. Similarly, *BDNF* is also involved in stress response and fear memory consolidation, as well as tameness during domestication, through its role in neural crest development. Based on these results, we propose that the strict association with humans in this species is linked with the evolution of pathways suppressing fear response and promoting tameness that are typically under selection in domesticated taxa.

Methods

Genome sequencing, assembly and annotation. HMW (High Molecular Weight) DNA was extracted from muscle tissue of a female barn swallow captured in a farm near Milan (Italy) and sequenced using 10x Genomics and Arima Hi-C technologies (Additional file 1: Supplementary Methods). Genomescope2.0 (43) was run online (http://qb.cshl.edu/genomescope/genomescope2.0/) starting from the *k*-mer (31 bp) histogram resulting from Meryl (115) (Additional file 1: Supplementary Methods). Newly generated data were combined with PacBio CLR long reads and Bionano optical maps already available for the same individual (31), using the VGP standard genome assembly pipeline 1.6 (36) (Additional file 1: Figure S1, Supplementary Methods). Briefly, Pacbio CLR long reads were assembled using FALCON (116), contigs were phased with FALCON-unzip (117) and polished with Arrow (smrtanalysis 5.1.0.26412). Two sets of contigs were generated, primary, representing one of the haplotypes, and alternate, representing the secondary haplotype. The primary contigs were purged (118), generating purged contigs and alternate haplotigs. The latter were merged with the alternate contigs and purged again. The primary purged contigs were then subjected to three steps of scaffolding with 10x linked reads, Bionano optical maps and Hi-C reads, generating chromosome-level scaffolds. Final scaffolds were merged with the alternate contigs and the mitogenome, generated with NOVOplasty (119) (Additional file 1: Supplementary Methods), polished with Arrow (smrtanalysis 5.1.0.26412) and Freebayes (120), and separated again in the two haplotypes, which then went through two steps of manual curation (121) (Additional file 1: Supplementary Methods). The primary curated assembly was annotated with IsoSeq and RNAseq data (Additional file 1: Table S7, Supplementary Methods).

Karyotype reconstruction. To confirm the chromosomal structure of our assembly, a karyotype for the barn swallow was generated using a cultured cell protocol (Additional file 1: Supplementary Methods). Chromosome sizes were predicted from the karyotype images and correlated with the assembly chromosome sizes (Additional file 1: Supplementary Methods).

Assembly evaluation and comparison with Chelidonia. Summary assembly statistics were computed with a script included in the VGP assembly pipeline GitHub repository (https://github.com/VGP/vgp-assembly/blob/master/pipeline/stats/asm_stats.sh). The assembly was further evaluated using BUSCO (44,122), Merqury (115), and Hi-C contact heatmaps (Additional file 1: Supplementary Methods). PacBio CLR long reads were aligned to the assembly and repeats were masked with a combination of Windowmasker

(123) and RepeatMasker (124,125) (Additional file 1: Supplementary Methods). The same procedure was applied to Chelidonia (31). A purge_dups (118) run was performed on the latter with default parameters. Correlations between bHirRus1 chromosome size and genomic content were performed with Spearman nonparametric rank tests (126) and Wilcoxon signed-rank tests (127) (Additional file 1: Supplementary Methods).

Cactus alignment. Progressive Cactus (50) v1.3.0 with default parameters was used initially to align bHirRus1 with 8 chromosome-level annotated Passeriformes genomes available on NCBI and the Chicken genome (Table S10). A maximum of 10 species were chosen for computational limits using Cactus. Despite different runs with the same parameters, two species failed to align (*Parus major* and *Ficedula albicollis*) and were excluded from the subsequent analyses. The guide tree was taken from TimeTree (128) (Additional file 1: Figure S7, Supplementary Methods). The genomes were soft-masked with WindowMasker (123) and RepeatMasker (124) (http://www.repeatmasker.org) (50) and then aligned (Additional file 1: Supplementary Methods). The alignment coverage was calculated with halAlignmentDepth (129) with the --noAncestors option and bHirRus1 as target species.

Neutral model estimation. PHAST v1.5 (130) was used in combination with the HAL toolkit (129) for the selection analyses. An alignment in the MAF format was extracted for each bHirRus1 chromosome from the Cactus HAL output using hal2maf (129) with the --noAncestors and --onlyOrthologs options. The MAFs were post-processed with maf_stream (https://github.com/joelarmstrong/maf_stream), as previously described (57). The non-conserved neutral model was trained from fourfold degenerate (4d) sites in the coding regions of the barn swallow annotation (55,131). Briefly, CDS that fall within bHirRus1 chromosomes were extracted from the NCBI gff3 annotation file. msa_view (130) 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 (--subst-mod REV --EM) to generate the neutral model.

PhyloP analysis. To detect conserved and accelerated bases, phyloP (130) was run on each chromosome separately using the neutral model with LRT method and in the CONACC mode. Due to the low total branch length between the aligned species (57), no significant calls were found after the false discovery rate (FDR) (52) 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 (126) was used to correlate chromosome size and the fraction covered by phyloP sites

(Additional file 1: Table S3). Wilcoxon signed-rank test (127) was used to compare differences between microchromosomes and the other chromosomes.

PhastCons analysis. An additional conservation analysis was performed using PhastCons (130) 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 (130). Each chromosome MAF file was split in chunks and 200 of them were randomly selected. phastCons was run on each chunk with initial parameters (Additional file 1: Supplementary Methods) generating tuned conserved and non-conserved models. These models were then used with phastCons to predict conservation scores and discrete conserved elements. Levels of smoothing and coverage were checked and the analysis was repeated again until the desired tuning was reached (Additional file 1: Supplementary Methods). Following Craig et al. (54), 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 (126).

Candidate genes detection and *CAMK2N2* tree construction. To detect candidate genes, we intersected the conserved and accelerated bases detected with each annotated class extracted with GenomicFeatures (Additional file 1: Supplementary Methods). Bases overlapping with more than one feature were assigned hierarchically based on the first appearance (54,132) in this order: CDS, 5' UTR, 3' UTR, intronic, intergenic. Genes without identified orthologs ("LOC" genes) were discarded. 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 and aligned with Muscle on MEGA (133). The tree was then generated using the Maximum likelihood method, a generalised time reversible (GTR) model and a gamma distribution (G) with 5 categories.

Gene ontology enrichment analysis. The gene ontology analysis was performed on the top 500 genes with the most overlaps with phyloP accelerated and conserved sites using *gage (134)* R package (Additional file 1: Supplementary Methods).

Generation of the pangenome and orthologs analysis. For the generation of the pangenome, additional 5 Italian barn swallow individuals were sampled (Additional file 1: Supplementary Methods). HMW DNA was extracted from the blood samples and sequenced with the PacBio HiFi technology (Additional file 1: Supplementary Methods). HiFi reads were checked for adaptor contamination and trimmed accordingly with cutadapt v3.2 (135) (Additional file 1: Supplementary Methods). Hifiasm (82) was used to assemble both primary and alternate assemblies which were then

purged with purge_dups (118) using custom cutoffs (83) (Table 31). Both primary and alternate HiFi-based assemblies were masked with Windowmasker (123) and RepeatMasker (124). The pangenome was generated with the Cactus (50) v1.3.0 Pangenome Pipeline (https://github.com/ComparativeGenomicsToolkit/cactus/blob/master/doc/pangenome.md, Additional file 1: Supplementary Methods) using HiFi-based assemblies and bHirRus1 primary and alternate assemblies. Orthologous genes were found running HALPER (136) 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 (129). Then, a file for the ortholog extension was generated from the coverage file and halLiftover (129) and used to lift bHirRus1 gene coordinates on the aligned HiFi assemblies. Orthologs were then found using the lifted genes. The resulting lists of orthologs were manually evaluated to find genes shared between individuals. The 236 genes that were found only in the bHirRus1 assembly were searched in the HiFi raw reads with BLAST 2.10.1+ (84). 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.

SNP catalogue. To generate the catalogue of genetic variants, all publicly available datasets were combined with our newly generated Hifi reads set (see "HiFi reads processing for genetic variants identification" Methods section). For each publicly available dataset, sequencing adapters and low quality bases were trimmed when present, and reads were aligned to the bHirRus1 reference genome. Freebayes v1.3 (120) was used on the alignments to call variants, parallelizing the process with a script from the VGP assembly pipeline (Additional file 1: Supplementary Methods). Variants were then split by population and markers were filtered for quality, read depth supporting each variant call, average fraction of missing sites among individuals and minor allele frequency (maf). Samples showing > 70% of missing genotypes were removed. Variants within repetitive regions were excluded, and only SNPs were extracted for downstream analysis. Details relative to the filters and threshold values used can be found in the Additional file 1: Supplementary Methods section.

For SNP density plotting and its correlation with genomic features, all data using the same sequencing technology were merged (HiFi WGS; Illumina WGS; Illumina ddRAD). SNP density was computed across all chromosomes (excluding unlocalized/unplaced scaffolds) over 10 kbp windows and these values were correlated with the GC content per window using the Spearman nonparametric rank test (126). SNPs falling in genic, intergenic, exonic and intronic regions (as determined from NCBI annotation) for each chromosome in the different datasets were counted. To plot SNP density

across all chromosomes, the KaryoploteR package (137) was used, computing its value over 40 kbp intervals (Additional file 1: Supplementary Methods).

Linkage disequilibrium analysis. Genome-wide LD decay was evaluated in all Illumina WGS datasets by calculating the r^2 coefficient using Plink v1.9 (138), considering all marker pairs within a 55 kbp distance. To calculate average r^2, SNP pairs were grouped according to their distance in bins of 1 kbp (range 1-55 kbp; Additional file 1: Supplementary Methods). 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 (Additional file 1: Supplementary Methods).

LD scans and extended haplotype homozygosity statistics. To scan chromosomes for regions containing alleles exhibiting high local LD values, Plink v1.9 (138) was used, considering marker pairs within a 15 kbp distance maximum. For the first LD scan, Illumina WGS data from ds3.1 were used. Each chromosome was divided into non-overlapping 5 kbp sliding windows to compute average LD (Additional file 1: Supplementary Methods). Next, only genomic windows with average $r^2 > 0.3$ were extracted and intersected with annotated features to generate a list of top candidate genes carrying alleles with high LD. Windows were excluded if in proximity (within ~5 kbp) with 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). Before computing within population haplotype homozygosity statistics (iHS) in ds3.1, ds2.1 and ds2.2, variants present on chr6 were phased and specifically filtered according to genotype missingness and maf parameters (Additional file 1: Supplementary Methods).

HiFi reads processing for genetic variants identification. HiFi reads from ds1 were aligned to bHirRus1 and small variants were called using deepvariant v1.0.0 (139). Only biallelic SNPs were kept, and variants falling within repetitive regions were removed. Next, variants were filtered according to genotype quality (quality > 20) and variant site depth (5% and 95% quantiles of the read depth values distribution were used to set the minimum and maximum site coverage). Joint variant calling of single-nucleotide variants (SNVs) and small insertions-deletions (indels) was performed using gVCF files from DeepVariant v1.1.0 per-sample calls, jointly called with GLNexus (140) pipeline (Additional file 1: Supplementary Methods). For structural variants (SVs), *pbsv* v2.6.0 (141) (commit v2.4.1-155-g281bd17) was used for per-sample and joint variant calling.

Titration and phasing experiments with HiFi reads. HiFi reads were first randomly downsampled and two titration experiments were conducted, the first one using variants obtained with individual variant calling and the second one with joint variant calling (N = 5). Estimation of haplotype-phased blocks length was also performed (Additional file 1: Supplementary Methods).

Data availability

Scripts used in this paper are available on GitHub (https://github.com/SwallowGenomics/BarnSwallow). Primary and alternate assemblies used in this study are available on NCBI under accession numbers GCF_015227805.1 and GCA_015227805.3. All raw data supporting the genome assembly are available on GenomeArk (https://vgp.github.io/genomeark/Hirundo_rustica/), and will also be available upon publication in SRA. The HiFi data will be made available upon publication. IsoSeq and RNAseq data are available on NCBI under the accession numbers SRR13516425, SRR13516426, SRR13516427, SRR9184408 and SRR9184409. The SNPs catalogue will be available upon publication on Dryad.

Acknowledgments

This work would have not been possible without the dedication of Prof. Nicola Saino. 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.); 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 Intramural Research Program of the National Library of Medicine, National Institutes of Health. We thank the INDACO Platform team (a project of High Performance Computing at the University of Milan, http://www.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) for euthanising and dissecting the barn swallow individual used for the assembly and annotation of bHirRus1 reference genome. We thank Dr. Alessandra Costanzo for her help in obtaining barn swallows blood samples.

Declaration of conflicts of interest

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

Author contributions

S.S., G.R.G., A.I, E.G, J.B, M.C., J.M, M.Sa., 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.So., A.I., J.F.O., R.S., P.M., K.W., L.G. and G.F. analysed the data.

S.S., G.R.G., M.So., A.B.-A., L.G. and G.F. drafted the manuscript.

C.C., A.P.M, T.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.

References

1. Johnston RF. Synanthropic birds of North America. In: Marzluff JM, Bowman R, Donnelly R, editors. Avian Ecology and Conservation in an Urbanizing World. Boston, MA: Springer US; 2001. p. 49–67.

2. Krajcarz M, Krajcarz MT, Baca M, Baumann C, Van Neer W, Popović D, et al. Ancestors of domestic cats in Neolithic Central Europe: Isotopic evidence of a synanthropic diet. Proc Natl Acad Sci U S A. 2020 Jul 28;117 (30):17710–9.

3. Ericsson M, Jensen P. Domestication and ontogeny effects on the stress response in young chickens (Gallus gallus). Sci Rep. 2016 Oct 26;6:35818.

4. Gogoleva SS, Volodin IA, Volodina EV, Kharlamova AV, Trut LN. Explosive vocal activity for attracting human attention is related to domestication in silver fox. Behav Processes. 2011 Feb;86 (2):216–21.

5. Ghazanfar AA, Kelly LM, Takahashi DY, Winters S, Terrett R, Higham JP. Domestication Phenotype Linked to Vocal Behavior in Marmoset Monkeys. Curr Biol. 2020 Dec 21;30 (24):5026–32.e3.

6. Wilkins AS, Wrangham RW, Fitch WT. The "domestication syndrome" in mammals: a unified explanation based on neural crest cell behavior and genetics. Genetics. 2014 Jul;197 (3):795–808.

7. Sánchez-Villagra MR, Geiger M, Schneider RA. The taming of the neural crest: a developmental perspective on the origins of morphological covariation in domesticated mammals. R Soc Open Sci. 2016 Jun;3 (6):160107.

8. Løtvedt P, Fallahshahroudi A, Bektic L, Altimiras J, Jensen P. Chicken domestication changes expression of stress-related genes in brain, pituitary and adrenals. Neurobiol Stress. 2017 Dec;7:113–21.

9. O'Rourke T, Martins PT, Asano R, Tachibana RO, Okanoya K, Boeckx C. Capturing the Effects of Domestication on Vocal Learning Complexity: (Trends in Cognitive Sciences 25, 462-474; 2021). Trends Cogn Sci. 2021 Aug;25 (8):722.

10. Hulme-Beaman A, Dobney K, Cucchi T, Searle JB. An Ecological and Evolutionary Framework for Commensalism in Anthropogenic Environments. Trends Ecol Evol. 2016 Aug;31 (8):633–45.

11. Harris SE, O'Neill RJ, Munshi-South J. Transcriptome resources for the white-footed mouse (Peromyscus leucopus): new genomic tools for investigating ecologically divergent urban and rural populations. Mol Ecol Resour. 2015 Mar;15 (2):382–94.

12. Brady SP. Road to evolution? Local adaptation to road adjacency in an amphibian (Ambystoma maculatum). Sci Rep. 2012 Jan 26;2:235.

13. Räsänen K, Laurila A, Merilä J. Geographic variation in acid stress tolerance of the moor frog, Rana arvalis. I. Local adaptation. Evolution. 2003 Feb;57 (2):352–62.

14. Whitehead A, Triant DA, Champlin D, Nacci D. Comparative transcriptomics implicates mechanisms of evolved pollution tolerance in a killifish population. Mol Ecol. 2010 Dec;19 (23):5186–203.

15. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013 Mar 21;495 (7441):360–4.

16. Hare B, Wobber V, Wrangham R. The self-domestication hypothesis: evolution of bonobo psychology is

due to selection against aggression. Anim Behav. 2012 Mar 1;83 (3):573–85.

17. Pap PL, Osváth G, Aparicio JM, Bărbos L, Matyjasiak P, Rubolini D, et al. Sexual Dimorphism and Population Differences in Structural Properties of Barn Swallow (Hirundo rustica) Wing and Tail Feathers. PLoS One. 2015 Jun 25;10 (6):e0130844.

18. Pap PL, Fülöp A, Adamkova M, Cepak J, Michalkova R, Safran RJ, et al. Selection on multiple sexual signals in two Central and Eastern European populations of the barn swallow. Ecol Evol. 2019 Oct;9 (19):11277–87.

19. Saino N, Romano M, Rubolini D, Ambrosini R, Romano A, Caprioli M, et al. A trade-off between reproduction and feather growth in the barn swallow (Hirundo rustica). PLoS One. 2014 May 14;9 (5):e96428.

20. Saino N, Ambrosini R, Albetti B, Caprioli M, De Giorgio B, Gatti E, et al. Migration phenology and breeding success are predicted by methylation of a photoperiodic gene in the barn swallow. Sci Rep. 2017 Mar 31;7:45412.

21. Saino N, Ambrosini R, Caprioli M, Liechti F, Romano A, Rubolini D, et al. Wing morphology, winter ecology, and fecundity selection: evidence for sex-dependence in barn swallows (Hirundo rustica). Oecologia. 2017 Aug;184 (4):799–812.

22. The Barn Swallow [Internet]. 2010. Available from: http://dx.doi.org/10.5040/9781472596888

23. Spina F. The EURING swallow project: a large-scale approach to the study and conservation of a long-distance migrant. In: Migrating birds know no boundaries Proceedings of the international symposium Israel: The Torgos. 1998. p. 151–62.

24. Safran RJ, Scordato ESC, Wilkins MR, Hubbard JK, Jenkins BR, Albrecht T, et al. Genome-wide differentiation in closely related populations: the roles of selection and geographic isolation. Mol Ecol. 2016 Aug;25 (16):3865–83.

25. von Rönn JAC, Shafer ABA, Wolf JBW. Disruptive selection without genome-wide evolution across a migratory divide. Mol Ecol. 2016 Jun;25 (11):2529–41.

26. Wilkins MR, Karaardıç H, Vortman Y, Parchman TL, Albrecht T, Petrželková A, et al. Phenotypic differentiation is associated with divergent sexual selection among closely related barn swallow populations. J Evol Biol. 2016 Dec;29 (12):2410–21.

27. Scordato ESC, Wilkins MR, Semenov G, Rubtsov AS, Kane NC, Safran RJ. Genomic variation across two barn swallow hybrid zones reveals traits associated with divergence in sympatry and allopatry. Mol Ecol. 2017 Oct;26 (20):5676–91.

28. Smith CCR, Flaxman SM, Scordato ESC, Kane NC, Hund AK, Sheta BM, et al. Demographic inference in barn swallows using whole-genome data shows signal for bottleneck and subspecies differentiation during the Holocene. Mol Ecol. 2018;27 (21):4200–12.

29. Schield DR, Scordato ESC, Smith CCR, Carter JK, Cherkaoui SI, Gombobaatar S, et al. Sex-linked genetic diversity and differentiation in a globally distributed avian species complex. Mol Ecol. 2021 May;30 (10):2313–32.

30. Zink RM, Pavlova A, Rohwer S, Drovetski SV. Barn swallows before barns: population histories and intercontinental colonization. Proc Biol Sci. 2006 May 22;273 (1591):1245–51.

31. Formenti G, Chiara M, Poveda L, Francoijs K-J, Bonisoli-Alquati A, Canova L, et al. 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 [Internet]. 2019 Jan 1;8 (1). Available from: http://dx.doi.org/10.1093/gigascience/giy142

32. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 2012 Aug 5;13:375.

33. Howe K, Wood JMD. Using optical mapping data for the improvement of vertebrate genome assemblies. Gigascience. 2015 Mar 18;4:10.

34. Zhang G, Rahbek C, Graves GR, Lei F, Jarvis ED, Gilbert MTP. Genomics: Bird sequencing project takes off. Nature. 2015 Jun 4;522 (7554):34.

35. Jarvis ED. Perspectives from the Avian Phylogenomics Project: Questions that Can Be Answered with Sequencing All Genomes of a Vertebrate Class. Annu Rev Anim Biosci. 2016;4:45–59.

36. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2020 May 23; (592):2020.05.22.110833.

37. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. Brief Bioinform. 2018 Jan 1;19 (1):118–35.

38. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. Nat Rev Genet. 2020 Apr;21 (4):243–54.

39. Wang H, Sawai A, Toji N, Sugioka R, Shibata Y, Suzuki Y, et al. Transcriptional regulatory divergence underpinning species-specific learned vocalization in songbirds. PLoS Biol. 2019 Nov;17 (11):e3000476.

40. Tomaszkiewicz M, Medvedev P, Makova KD. Y and W Chromosome Assemblies: Approaches and Discoveries. Trends Genet. 2017 Apr;33 (4):266–82.

41. Malinovskaya LP, Tishakova K, Shnaider EP, Borodin PM, Torgasheva AA. Heterochiasmy and Sexual Dimorphism: The Case of the Barn Swallow (Hirundo rustica, Hirundinidae, Aves). Genes [Internet]. 2020 Sep 24;11 (10). Available from: http://dx.doi.org/10.3390/genes11101119

42. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004 Dec 9;432 (7018):695–716.

43. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020 Mar 18;11 (1):1432.

44. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol. 2018 Mar 1;35 (3):543–8.

45. Kuhl H, Frankl-Vilches C, Bakker A. An unbiased molecular approach using 3′-UTRs resolves the avian family-level tree of life. Mol Biol [Internet]. 2021; Available from: https://academic.oup.com/mbe/article-abstract/38/1/108/5891114

46. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014 Jan;42 (Database issue):D756–63.

47. Francis WR, Wörheide G. Similar Ratios of Introns to Intergenic Sequence across Animal Genomes.

Genome Biol Evol. 2017 Jun 1;9 (6):1582–98.

48. Burt DW. Origin and evolution of avian microchromosomes. Cytogenet Genome Res. 2002;96 (1-4):97–112.

49. Kim J, Lee C, Ko BJ, Yoo DA, Won S, Phillippy A. False gene and chromosome losses affected by assembly and sequence errors. bioRxiv [Internet]. 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.04.09.438906v1.abstract

50. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020 Nov;587 (7833):246–51.

51. Armstrong J. Enabling comparative genomics at the scale of hundreds of species [Internet]. UC Santa Cruz; 2019 [cited 2021 Mar 5]. Available from: https://escholarship.org/uc/item/7pv8w2bz

52. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing [Internet]. Vol. 57, Journal of the Royal Statistical Society: Series B (Methodological). 1995. p. 289–300. Available from: http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x

53. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014 Dec 12;346 (6215):1311–20.

54. Craig RJ, Suh A, Wang M, Ellegren H. Natural selection beyond genes: Identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (Ficedula albicollis). Mol Ecol. 2018 Jan;27 (2):476–92.

55. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug;15 (8):1034–50.

56. Axelsson E, Webster MT, Smith NGC, Burt DW, Ellegren H. Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. Genome Res. 2005 Jan;15 (1):120–5.

57. Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, et al. Dense sampling of bird diversity increases power of comparative genomics. Nature. 2020 Nov;587 (7833):252–7.

58. Malinow R, Schulman H, Tsien RW. Inhibition of postsynaptic PKC or CaMKII blocks induction but not expression of LTP. Science. 1989 Aug 25;245 (4920):862–6.

59. Silva AJ, Stevens CF, Tonegawa S, Wang Y. Deficient hippocampal long-term potentiation in alpha-calcium-calmodulin kinase II mutant mice. Science. 1992 Jul 10;257 (5067):201–6.

60. Hayashi Y, Shi SH, Esteban JA, Piccini A, Poncer JC, Malinow R. Driving AMPA receptors into synapses by LTP and CaMKII: requirement for GluR1 and PDZ domain interaction. Science. 2000 Mar 24;287 (5461):2262–7.

61. Benke TA, Lüthi A, Isaac JT, Collingridge GL. Modulation of AMPA receptor unitary conductance by synaptic activity. Nature. 1998 Jun 25;393 (6687):793–7.

62. Derkach V, Barria A, Soderling TR. Ca2+/calmodulin-kinase II enhances channel conductance of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionate type glutamate receptors. Proc Natl Acad Sci U S A. 1999 Mar 16;96 (6):3269–74.

63. Giese KP, Mizuno K. The roles of protein kinases in learning and memory. Learn Mem. 2013 Sep 16;20

(10):540–52.

64. Buard I, Coultrap SJ, Freund RK, Lee Y-S, Dell'Acqua ML, Silva AJ, et al. CaMKII "Autonomy" Is Required for Initiating But Not for Maintaining Neuronal Long-Term Information Storage [Internet]. Vol. 30, Journal of Neuroscience. 2010. p. 8214–20. Available from: http://dx.doi.org/10.1523/jneurosci.1469-10.2010

65. Vigil FA, Mizuno K, Lucchesi W, Valls-Comamala V, Giese KP. Prevention of long-term memory loss after retrieval by an endogenous CaMKII inhibitor. Sci Rep. 2017 Jun 22;7 (1):4040.

66. Satterlee DG, Johnson WA. Selection of Japanese quail for contrasting blood corticosterone response to immobilization. Poult Sci. 1988 Jan;67 (1):25–32.

67. Khatri B, Kang S, Shouse S, Anthony N, Kuenzel W, Kong BC. Copy number variation study in Japanese quail associated with stress related traits using whole genome re-sequencing data. PLoS One. 2019 Mar 28;14 (3):e0214543.

68. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature. 2001 Oct 4;413 (6855):519–23.

69. Fisher SE, Scharff C. FOXP2 as a molecular window into speech and language. Trends Genet. 2009 Apr;25 (4):166–77.

70. Bacon C, Rappold GA. The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders. Hum Genet. 2012 Nov;131 (11):1687–98.

71. Sollis E, Graham SA, Vino A, Froehlich H, Vreeburg M, Dimitropoulou D, et al. Identification and functional characterization of de novo FOXP1 variants provides novel insights into the etiology of neurodevelopmental disorder. Hum Mol Genet. 2016 Feb 1;25 (3):546–57.

72. Siper PM, De Rubeis S, Trelles MDP, Durkin A, Di Marino D, Muratet F, et al. Prospective investigation of FOXP1 syndrome. Mol Autism. 2017 Oct 24;8:57.

73. Morgan AT, Webster R. Aetiology of childhood apraxia of speech: A clinical practice update for paediatricians. J Paediatr Child Health. 2018 Oct;54 (10):1090–5.

74. Teramitsu I, Kudo LC, London SE, Geschwind DH, White SA. Parallel FoxP1 and FoxP2 expression in songbird and human brain predicts functional interaction. J Neurosci. 2004 Mar 31;24 (13):3152–63.

75. Mendoza E, Tokarev K, Düring DN, Retamosa EC, Weiss M, Arpenik N, et al. Differential coexpression of FoxP1, FoxP2, and FoxP4 in the Zebra Finch (Taeniopygia guttata) song system. J Comp Neurol. 2015 Jun 15;523 (9):1318–40.

76. Norton P, Barschke P, Scharff C, Mendoza E. Differential Song Deficits after Lentivirus-Mediated Knockdown of FoxP1, FoxP2, or FoxP4 in Area X of Juvenile Zebra Finches. J Neurosci. 2019 Dec 4;39 (49):9782–96.

77. Garcia-Oscos F, Koch T, Pancholi H, Trusel M, Daliparthi V, Ayhan F, et al. Autism-linked gene FoxP1 selectively regulates the cultural transmission of learned vocalizations. Sci Adv [Internet]. 2021;7 (6). Available from: https://www.science.org/doi/full/10.1126/sciadv.abd2827

78. Oikkonen J, Huang Y, Onkamo P, Ukkola-Vuoti L, Raijas P, Karma K, et al. A genome-wide linkage and association study of musical aptitude identifies loci containing genes related to inner ear development and neurocognitive functions. Mol Psychiatry. 2015 Feb;20 (2):275–82.

79. Park H, Lee S, Kim H-J, Ju YS, Shin J-Y, Hong D, et al. Comprehensive genomic analyses associate

UGT8 variants with musical ability in a Mongolian population. J Med Genet. 2012 Dec;49 (12):747–52.

80. Oikkonen J, Kuusi T, Peltonen P, Raijas P, Ukkola-Vuoti L, Karma K, et al. Creative Activities in Music – A Genome-Wide Linkage Analysis [Internet]. Vol. 11, PLOS ONE. 2016. p. e0148679. Available from: http://dx.doi.org/10.1371/journal.pone.0148679

81. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome Graphs. Annu Rev Genomics Hum Genet. 2020 Aug 31;21:139–62.

82. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021 Feb;18 (2):170–5.

83. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020 Sep;30 (9):1291–305.

84. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421.

85. Smeds L, Warmuth V, Bolivar P, Uebbing S, Burri R, Suh A, et al. Evolutionary analysis of the female-specific avian W chromosome. Nat Commun. 2015 Jun 4;6:7330.

86. Zhao Z, Fu Y-X, Hewett-Emmett D, Boerwinkle E. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene. 2003 Jul 17;312:207–13.

87. Makova KD, Hardison RC. The effects of chromatin organization on variation in mutation rates in the genome. Nat Rev Genet. 2015 Apr;16 (4):213–23.

88. Kim Y, Nielsen R. Linkage disequilibrium as a signature of selective sweeps. Genetics. 2004 Jul;167 (3):1513–24.

89. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006 Mar;4 (3):e72.

90. Ennis S. Linkage disequilibrium as a tool for detecting signatures of natural selection. Methods Mol Biol. 2007;376:59–70.

91. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science. 2008 Mar 7;319 (5868):1395–8.

92. Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, et al. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. Mol Ecol. 2017 Aug;26 (16):4158–72.

93. Zaitlen N, Huntsman S, Hu D, Spear M, Eng C, Oh SS, et al. The Effects of Migration and Assortative Mating on Admixture Linkage Disequilibrium. Genetics. 2017 Jan;205 (1):375–83.

94. Bürger R, Akerman A. The effects of linkage and gene flow on local adaptation: a two-locus continent-island model. Theor Popul Biol. 2011 Dec;80 (4):272–88.

95. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002 Jun 21;296 (5576):2225–9.

96. Zhou Y, Qiu H, Xu S. Modeling Continuous Admixture Using Admixture-Induced Linkage

Disequilibrium. Sci Rep. 2017 Feb 23;7:43054.

97. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008 Jun;9 (6):477–85.

98. Joiret M, Mahachie John JM, Gusareva ES, Van Steen K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. BioData Min. 2019 Jun 10;12:11.

99. Liu S, He S, Chen L, Li W, Di J, Liu M. Estimates of linkage disequilibrium and effective population sizes in Chinese Merino (Xinjiang type) sheep by genome-wide SNPs. Genes Genomics. 2017 Apr 17;39 (7):733–45.

100. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000 Jul;67 (1):170–81.

101. Stapley J, Birkhead TR, Burke T, Slate J. Pronounced inter- and intrachromosomal variation in linkage disequilibrium across the zebra finch genome. Genome Res. 2010 Apr;20 (4):496–502.

102. Kapusta A, Suh A. Evolution of bird genomes-a transposon's-eye view. Ann N Y Acad Sci. 2017 Feb;1389 (1):164–85.

103. Monteggia LM, Barrot M, Powell CM, Berton O, Galanis V, Gemelli T, et al. Essential role of brain-derived neurotrophic factor in adult hippocampal function. Proc Natl Acad Sci U S A. 2004 Jul 20;101 (29):10827–32.

104. Bramham CR, Messaoudi E. BDNF function in adult synaptic plasticity: the synaptic consolidation hypothesis. Prog Neurobiol. 2005 Jun;76 (2):99–125.

105. Li XC, Jarvis ED, Alvarez-Borda B, Lim DA, Nottebohm F. A relationship between behavior, neurotrophin expression, and new neuron survival. Proc Natl Acad Sci U S A. 2000 Jul 18;97 (15):8584–9.

106. Dittrich F, Feng Y, Metzdorf R, Gahr M. Estrogen-inducible, sex-specific expression of brain-derived neurotrophic factor mRNA in a forebrain song control nucleus of the juvenile zebra finch. Proc Natl Acad Sci U S A. 1999 Jul 6;96 (14):8241–6.

107. Dittrich F, Ter Maat A, Jansen RF, Pieneman A, Hertel M, Frankl-Vilches C, et al. Maximized song learning of juvenile male zebra finches following BDNF expression in the HVC. Eur J Neurosci. 2013 Nov;38 (9):3338–44.

108. Sieber-Blum M. Role of the neurotrophic factors BDNF and NGF in the commitment of pluripotent neural crest cells. Neuron. 1991 Jun;6 (6):949–55.

109. Notaras M, van den Buuse M. Neurobiology of BDNF in fear memory, sensitivity to stress, and stress-related disorders. Mol Psychiatry. 2020 Oct;25 (10):2251–74.

110. Maynard KR, Hill JL, Calcaterra NE, Palko ME, Kardian A, Paredes D, et al. Functional Role of BDNF Production from Unique Promoters in Aggression and Serotonin Signaling. Neuropsychopharmacology. 2016 Jul;41 (8):1943–55.

111. Yossifoff M, Kisliouk T, Meiri N. Dynamic changes in DNA methylation during thermal control establishment affect CREB binding to the brain-derived neurotrophic factor promoter. Eur J Neurosci. 2008 Dec;28 (11):2267–77.

112. George JM, Bell ZW, Condliffe D, Dohrer K, Abaurrea T, Spencer K, et al. Acute social isolation alters

neurogenomic state in songbird forebrain. Proc Natl Acad Sci U S A. 2020 Sep 22;117 (38):23311–6.

113. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005 Nov;15 (11):1566–75.

114. Lubin FD, Roth TL, Sweatt JD. Epigenetic regulation of BDNF gene transcription in the consolidation of fear memory. J Neurosci. 2008 Oct 15;28 (42):10576–86.

115. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245.

116. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013 Jun;10 (6):563–9.

117. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016 Dec;13 (12):1050–4.

118. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020 May 1;36 (9):2896–8.

119. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017 Feb 28;45 (4):e18.

120. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv [q-bio.GN]. 2012. Available from: http://arxiv.org/abs/1207.3907

121. Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, et al. Significantly improving the quality of genome assemblies through curation. Gigascience [Internet]. 2021 Jan 9;10 (1). Available from: http://dx.doi.org/10.1093/gigascience/giaa153

122. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31 (19):3210–2.

123. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2006 Jan 15;22 (2):134–41.

124. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009 Mar;Chapter 4:Unit 4.10.

125. Smit AFA. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0.

126. Zar JH. Significance Testing of the Spearman Rank Correlation Coefficient. J Am Stat Assoc. 1972 Sep 1;67 (339):578–80.

127. Wilcoxon F. Individual comparisons by ranking methods, Biometrics Bulletin 1 (1945) 80--83. URL: http://www jstor org/stable/3001968 doi. 10:3001968.

128. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017 Jul 1;34 (7):1812–9.

129. Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics. 2013 May 15;29 (10):1341–2.

130. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time

models [Internet]. Vol. 12, Briefings in Bioinformatics. 2011. p. 41–51. Available from: http://dx.doi.org/10.1093/bib/bbq072

131. Siepel A. PhastCons HOWTO. Available from: http://compgen.bscb.cornell.edu/phast/phastCons-HOWTO.html

132. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011 Oct 12;478 (7370):476–82.

133. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018 Jun 1;35 (6):1547–9.

134. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis [Internet]. Vol. 10, BMC Bioinformatics. 2009. Available from: http://dx.doi.org/10.1186/1471-2105-10-161

135. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2;17 (1):10–2.

136. Zhang X, Kaplow IM, Wirthlin M, Park TY, Pfenning AR. HALPER facilitates the identification of regulatory element orthologs across species. Bioinformatics. 2020 Aug 1;36 (15):4339–40.

137. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics. 2017 Oct 1;33 (19):3088–90.

138. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81 (3):559–75.

139. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018 Nov;36 (10):983–7.

140. Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics [Internet]. 2021 Jan 5; Available from: http://dx.doi.org/10.1093/bioinformatics/btaa1081

141. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019 Oct;37 (10):1155–62.

142. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019 Aug;15 (8):e1007273.