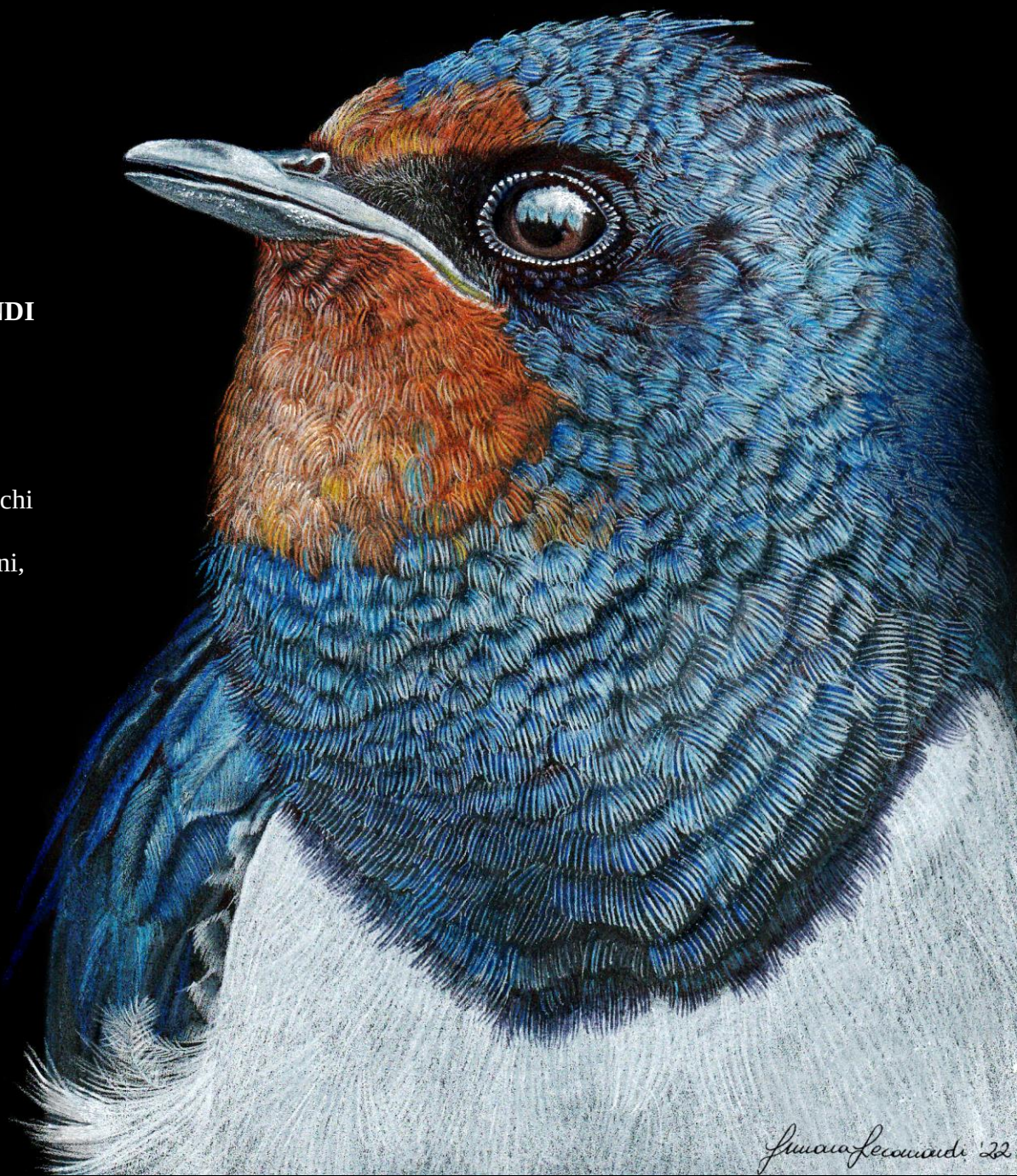
UNIVERSITÀ DEGLI STUDI DI MILANO
PhD Course in Environmental Sciences XXXIV Cycle
Department of Biosciences

CHROMOSOME-LEVEL *DE NOVO* GENOME ASSEMBLIES OF AVIAN SPECIES AND THEIR RELEVANCE FOR COMPARATIVE GENOMICS, PANGENOMICS, POPULATION GENOMICS AND SPECIES CONSERVATION



PhD Thesis
SIMONA SECOMANDI
R12425

Tutor:
Prof. Luca Gianfranceschi
Co-tutors:
Prof. Roberto Ambrosini,
Dott. Giulio Formenti

Head of the PhD course:
Prof. Francesco Ficetola

Academic Year: 2020-2021

This thesis is dedicated
to Nicola Saino.

Table of contents

ABSTRACT

English

Life on Earth is currently experiencing the sixth mass extinction. A major loss of species is being caused by anthropogenic habitat destruction, illegal wildlife trade, overfishing, massive fossil fuel consumption and the consequent climate changes, leading to a progressive collapse of biodiversity. The establishment of conservation initiatives is now more crucial than ever and the genetic management of threatened species is currently gaining critical significance. Previous genetic studies have been limited by the lack of complete reference genomes, thus having to focus on a limited number of genes or incomplete genomic sequencing data. In several cases, this led to wrong or incomplete assumptions and to conservation decisions which did not carry the desired effects due to incorrect evaluations. The availability of high-throughput sequencing technologies and the development of advanced computational methods have recently allowed the generation of cost-effective chromosome-level reference genomes. In the last decades, tremendous efforts were made to generate the most complete human genome. However, additional reference genomes spanning the entire tree of life are a necessary foundation for the study of biology in the 21st century, which was made possible with the establishment of large sequencing consortia. This thesis work represents a contribution to the collective effort of describing and preserving our planet's genetic diversity, being part of international consortia such as the Vertebrate Genomes Project (VGP), Darwin Tree of Life (DToL) and European Reference Genome Atlas (ERGA). Here I report new methods to assemble highly contiguous reference genomes for vertebrate species in the context of the VGP and their relevance in deciphering the biology of a species, its evolution, its intrinsic variability and in planning conservation actions, providing a comprehensive collection of genomic markers. Using the VGP pipelines, I have assembled 6 chromosome-level bird species: barn swallow (*Hirundo rustica*), lesser kestrel (*Falco naumanni*), American flamingo (*Phoenicopterus ruber*), common yellowthroat (*Geothlypis trichas*), rifleman (*Acanthisitta chloris*) and red-crested turaco (*Tauraco erythrolophus*), which fulfilled and exceeded the standard VGP metrics for genome assembly. Moreover, I contributed to the generation of the European nightjar (*Caprimulgus europaeus*) reference genome for the DToL. The genomes I generated will be exploited in the future to get insights into the biology of these species, but also for comparative genomics analyses using all available VGP-quality species. In particular, the lesser kestrel reference genome will be a fundamental resource for a future study that involves the assessment of how this species coped with climatic fluctuations in the past and how it is expected to cope with them in the future under the current scenario of climate changes. Moreover, the European nightjar genome will help to deepen the knowledge on the biology of this cryptic and

elusive bird, also boosting the sequencing of other members of the Caprimulgidae family to reconstruct a comprehensive phylogeny. During my PhD, I particularly focused on barn swallow (*Hirundo rustica*) genomics, an iconic migratory passerine bird with a long-standing association with humans. Using the VGP-quality assembly, I performed comparative genomics, population genomics and pangenomics analyses. Comparative genomics work was carried out with the reference-free aligner Cactus, with which I have aligned the barn swallow reference with other chromosome-level bird species. Using the alignment, I performed a positive and negative selection analysis across the genome to find candidate genes under selection important for barn swallow biology. For population genomics analyses, we aligned all publicly available data for the barn swallow subspecies and populations to the reference genome and performed a Linkage Disequilibrium scan. Comparative and population genomics approaches both pointed at candidate genes under selective constraint which may have a role in the onset of the synanthropic behavior of the species. Finally, with the reference genome and other HiFi-based barn swallow assemblies, I constructed the first pangenome graph for the species and preliminarily evaluated the extent of core and accessory genes among individuals. We also complemented the nuclear genome-based study with the generation of the complete mitochondrial genome for the species, which allowed us to dissect the phylogenetic relationships between barn swallow subspecies and to clarify the species' phylogeographic history.

In conclusion, this thesis work represents a valuable step forward and a contribution to international genome sequencing efforts. I outlined how complete genomic resources can revolutionize studies on the biology and evolution of a species, but also how they represent a pivotal resource to correctly plan threatened species conservation actions during the sixth mass extinction.

Italian

Il nostro pianeta sta attualmente affrontando la sesta estinzione di massa. La distruzione degli habitat, il bracconaggio, la pesca intensiva, il massiccio consumo di combustibili fossili e i conseguenti cambiamenti climatici, hanno provocato un'elevata perdita di specie che porterà ben presto al collasso della biodiversità. La fondazione di iniziative per la conservazione delle specie è ora più cruciale che mai e la loro gestione dal punto di vista genetico sta acquisendo una rilevanza sempre più critica. Studi precedenti sono stati limitati dalla mancanza di genomi di riferimento completi, dovendo focalizzarsi quindi su un numero limitato di geni o su risorse genetiche incomplete. Questo, in alcuni casi, ha portato ad ipotesi errate o incomplete, ma anche all'attuazione di piani di conservazione che non hanno apportato i benefici previsti a causa di errate valutazioni. La disponibilità di innovative tecniche di sequenziamento e lo sviluppo di nuovi metodi computazionali hanno recentemente consentito l'assemblaggio degli interi cromosomi che costituiscono il cariotipo di una specie. Negli ultimi decenni sono stati compiuti enormi sforzi per generare il genoma umano più completo possibile. Tuttavia, la disponibilità di ulteriori genomi di riferimento lungo tutto l'albero della vita è una base necessaria per la biologia del 21° secolo, cosa che è stata resa possibile dalla fondazione di consorzi internazionali per il sequenziamento di genomi. Questo lavoro di tesi rientra nel contesto di grandi consorzi di sequenziamento, quali Vertebrate Genomes Project (VGP), Darwin Tree of Life (DToL) e European Reference Genome Atlas (ERGA), rappresentando quindi un contributo allo sforzo collettivo di descrivere e preservare la biodiversità che caratterizza il nostro pianeta. La tesi riporta nuovi metodi per l'assemblaggio di genomi a livello cromosomico per il VGP e la loro rilevanza nel decifrare la biologia di una specie, la sua evoluzione, la sua variabilità intrinseca e la gestione dei piani di conservazione, fornendo infatti un catalogo completo di marcatori genetici che possono essere sfruttati nella pianificazione. Utilizzando le *pipeline* di assemblaggio del VGP, ho assemblato 6 specie di uccelli: rondine (*Hirundo rustica*), grillaio (*Falco naumanni*), fenicottero rosso (*Phoenicopterus ruber*), golagialla comune (*Geothlypis trichas*), fuciliere (*Acanthisitta chloris*) e turaco crestarossa (*Tauraco erythrolophus*), i quali hanno soddisfatto e superato le statistiche minime in termini di contiguità e qualità dei genomi imposte dal VGP. Inoltre, ho contribuito alla generazione del genoma di riferimento del succiacapre (*Caprimulgus europaeus*) per il DToL. I genomi che ho generato saranno utilizzati in studi futuri per approfondire la biologia di queste specie, ma anche per effettuare studi di genomica comparativa usando tutti i genomi VGP disponibili. In particolare, il genoma di riferimento del grillaio costituirà una risorsa fondamentale per determinare come questo piccolo falco migratore ha reagito alle fluttuazioni climatiche avvenute in passato e come reagirà in un futuro segnato dai cambiamenti climatici. Il genoma del succiacapre, invece, risulterà utile per approfondire la biologia di questa specie criptica ed elusiva, incoraggiando anche il sequenziamento di altre specie di Caprimulgidae e il delineamento della loro filogenesi. In questa tesi, mi sono

concentrata sulla genomica della rondine comune (*Hirundo rustica*), un iconico passeriforme migratore che vive a stretto contatto con l'uomo. Utilizzando il genoma che ho assemblato con la *pipeline* del VGP, ho eseguito analisi di genomica comparativa, genomica di popolazione e pangenomica. Nel lavoro di genomica comparativa ho usato il software Cactus, con il quale ho allineato il genoma della rondine con quelli di altre specie di uccelli, eseguendo poi un'analisi di selezione, sia positiva che negativa, lungo il genoma per trovare geni candidati importanti per la biologia della rondine. Per le analisi di genomica di popolazione, tutti i dati pubblici disponibili per popolazioni e sottospecie di rondine sono stati allineati al genoma di riferimento e sono stati misurati i livelli di Linkage Disequilibrium. Entrambi gli approcci hanno evidenziato una selezione per geni candidati che potrebbero avere avuto un ruolo nell'insorgenza dei comportamenti sinantropici della specie. Inoltre, con il genoma di riferimento e quelli di altri individui di rondine sequenziati con la tecnologia HiFi, ho generato il primo pangenoma per la specie e osservato preliminarmente i geni persi o acquisiti dai vari individui. La sezione basata sul genoma nucleare è stata inoltre integrata dalla generazione del genoma mitocondriale completo della rondine che ha consentito di delineare le relazioni filogenetiche tra le sue sottospecie e di chiarire la storia filogeografica della specie.

Per concludere, questo lavoro di tesi rappresenta un prezioso passo avanti e un contributo agli sforzi attuati dai consorzi internazionali di sequenziamento dei genomi. Ho sottolineato come l'utilizzo di risorse genomiche complete possa rivoluzionare gli studi sulla biologia e l'evoluzione delle specie, ma anche come esse rappresentino una risorsa fondamentale per pianificare correttamente le azioni di conservazione delle specie minacciate dalla sesta era di estinzione di massa.

INTRODUCTION

1. Overview of past and current DNA sequencing technologies

1.1 First-Generation Sequencing

In 1953, Francis Crick and James Watson discovered the double-helix structure of DNA[1]. Since then, discovering the exact genomic sequences of living organisms has become a major goal for humanity. In the two decades following Crick and Watson discovery, the first biological molecules were sequenced[2–10] (reviewed in Giani et al., 2020[11]). In particular, in 1975, Friederick Sanger was able to sequence the entire 5 kilo base pairs (kbp) long genome of Bacteriophage ΦX174 with its "plus and minus" sequencing method[12]. However, early methods required a large effort in terms of time and labor to sequence a few nucleotides[13]. It was only in 1977 that the first techniques for sequencing in a short time were developed[13]. In the original Sanger method[14], which could sequence up to 400 bp, there are four reactions each containing a single-stranded DNA template, a DNA primer, a DNA polymerase, 4 deoxynucleotide triphosphates (dNTPs) and a small fraction of a single di-deoxynucleotide triphosphate (ddNTPs), that, when incorporated, stop the reaction, generating DNA fragments of variable length. Allan Maxam and Walter Gilbert, on the other hand, devised a method based on the differential modification of the bases and subsequent cleavage of the DNA at the modified nucleotides[15]. In both methods, DNA fragments are separated by gel electrophoresis[16], X-ray films are used to detect the DNA fragments length and this information is used to ultimately reconstruct the original sequence. In 1979, the concept of "shotgun sequencing" was introduced to improve the speed at which DNA could be read[17]. In shotgun sequencing, the DNA is fragmented and each fragment is sequenced individually. The assembly of the short fragments is performed by exploiting the partial overlaps between sequences, potentially with the aid of a computer program, determining their position and order in the template sequence. In 1982, this method was used to sequence the entire bacteriophage Lambda genome[18] (48 Mega base pairs, Mbp, long) and, two years later, the 152 Mbp long Epstein-Barr virus B59-8 strain[19]. The development of the PCR (Polymerase Chain Reaction) in 1983 sped up the process even further, by increasing the available starting material amount with amplification[20]. In the following years, the original time-consuming Sanger method was refined and automated[21–25]. New implementations included the use of a single reaction instead of four, the switch from radioactive or fluorescent labeling and the improvement of the accuracy and length of the output sequences (up to 1,000 bp). In parallel, through the late 80s and early 90s a new method for DNA sequencing[26–28] called "pyrosequencing" was developed. Similarly to Sanger sequencing, pyrosequencing also relies on "sequencing by synthesis". However, in pyrosequencing the detection of nucleotide incorporation is achieved by measuring light emitted by pyrophosphate release during DNA polymerase cleavage using a firefly luciferase[29]. Major advantages include the use of

unmodified nucleotides and the opportunity to sequence in real-time[30]. Additional progress was made during the 90s, including wet laboratory protocols improvements and the introduction of "paired-end" sequencing (reviewed in Giani et al., 2020[11]). With this technique, both ends of a DNA fragment are sequenced, allowing to read longer DNA fragments, improving the base-calling accuracy and helping to resolve the assembly of repetitive regions[1]. Given the increasing availability of sequencing data, many sequencing projects were launched and online repositories, such as the DNA sequence database Genbank[31] in 1982, were created to collect all the newly generated data. Improvements in data analysis tools were also made, with the birth of bioinformatics in the 1980s[11].

1.2 Second-Generation Sequencing

In the early 2000, the advent of Next Generation Sequencing (NGS) technologies, now often referred to as Second Generation Sequencing (SGS), revolutionized the field. SGS methods parallelize sequencing reactions in mass, allowing to obtain a much greater amount of DNA sequences in a single run compared to early technologies[32]. SGS first step is library preparation, in which the DNA is fragmented, and fragments are size selected and labeled by attaching adapters at both ends[33,34]. The first SGS sequencer was produced by 454 Life Sciences, which implemented the pyrosequencing method for massive parallel sequencing[32]. Later on, Solexa launched the most used parallel sequencing technique, then acquired by Illumina[35]. In Illumina sequencing, a reduced cycle amplification adds adapters at the end of DNA fragments, which include sequencing binding sites, indexes and complementary regions to two types of oligos bound to a flow cell. On the flow cell surface, each fragment is isothermally amplified through a process called bridge amplification[36,37]. During bridge amplification, each DNA fragment binds to the first type of oligos with one of its modified ends. A DNA polymerase replicates the fragment using the 3'-end of the flow cell-bound oligo as primer. The obtained double stranded DNA is denatured and the original fragment is eliminated, while the newly synthesized strand arches over the flow cell and binds to the second type of oligos, creating a "bridge"[35]. The DNA polymerase, again, replicates the strand. The DNA is denatured and the two complementary strands are now bound to the flow cell individually. At the end of the amplification process, only the forward strand is maintained. This amplification generates clusters of sequences covalently bound to the flow cell. After bridge amplification, the "sequencing by synthesis" step begins with the incorporation of fluorescent-tagged nucleotides complementary to the sequence[38]. Each nucleotide acts as a reversible chain terminator, permitting to incorporate one dNTP at each cycle[39–41]. After each addition, a light source excites the added nucleotides in the cluster which emit a fluorescent signal. The signal intensity and wavelength determine the base call[2].

[1] DNA motifs that appear in multiple copies throughout a genome.
[2] See https://www.youtube.com/watch?v=fCd6B5HRaZ8

Although Illumina sequencing is currently by far the most widespread method to read DNA, over the years several competing technologies were developed. An alternative sequencing approach was SOLiD from Applied Biosystems, developed in 2007. This method relies on the ligation of single copy DNA templates onto microbeads which were then subjected to *in vitro* cloning[42]. Several cycles of ligation-based sequencing with fluorescent oligonucleotides are performed on millions of microbeads in parallel[42,43]. Another method, Ion Torrent[44], was developed in 2011. It is based on the detection of hydrogen ions released during the addition of a nucleotide. Its limitation was the incorrect estimation of homopolymers length[45,46].

SGS, although being a "short-read sequencing" (SRS) approach (75-400 bp)[47] yields to an accuracy of 99.9%[13]. However, sequencing short DNA fragments put several limitations in the assembly of large genomes, impacting the subsequent analysis. This is mostly because repetitive elements can not be completely resolved whenever they are longer than the read length, ultimately fragmenting the genome assemblies[48]. As it will be further discussed, this introduces several shortcomings for downstream analyses. For instance, contiguity is pivotal to accurately identify large structural variants (SVs) as insertions and deletions (indels), inversions, duplications and translocations, which have been demonstrated to play important biological roles[49,50]. With short reads it was possible to identify only 10%[51] to 70%[49,52] of total SVs content, with many false positives[49,53–56]. On the other hand, SGS allows several library preparations that can greatly benefit the genome assembly process (e.g. see section *1.2.1*). Two of these approaches that were employed in this PhD thesis work are outlined in the next two paragraphs.

1.2.1 10x Linked-Reads

In 2016, 10x Genomics released a new sequencing technology called Linked-Reads[57] (https://www.10xgenomics.com/products/linked-reads). With this technology, it was possible to retrieve contextual genomic information from short reads. During library preparation, High Molecular Weight (HMW) DNA is partitioned (> 50 kbp long fragments) and barcoded inside functionalized gel beads (Gel Bead-in-Emulsions, GEMs, **Figure 1**). GEMs undergo a step of isothermal incubation to generate 10x barcode amplicons that are sequenced with the standard Illumina method to generate barcoded paired-end short reads. Each fragment in the GEM has the same barcode since they come from the same long DNA molecule. Barcodes are useful to correctly place reads in hard-to-assemble regions, using barcoded anchors to recruit short reads into paralogous loci. Linked-Reads can also be useful to phase the genome (i.e. separate the two divergent haplotypes of an individual)[57], and detect structural variants[58].

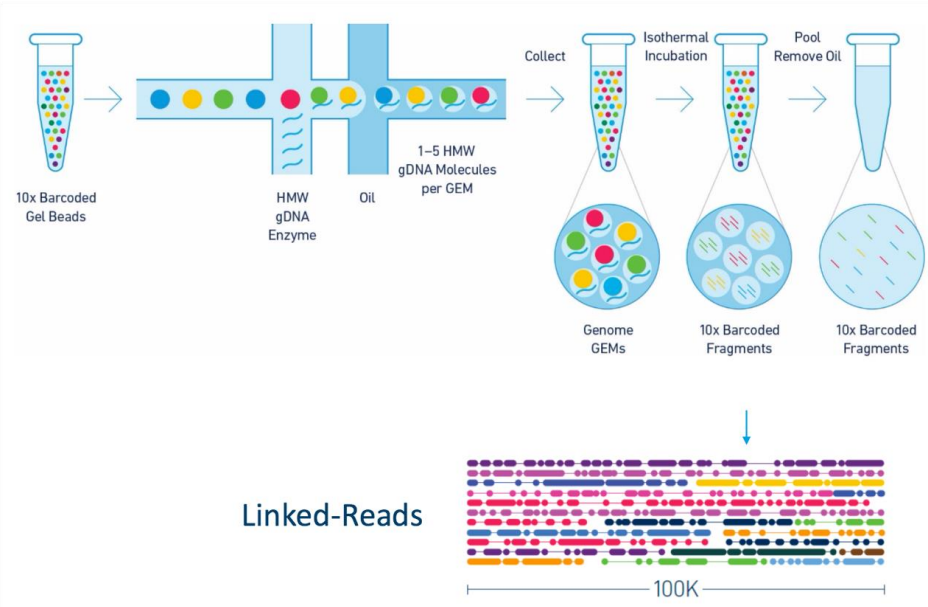


Figure 1. Overview of 10x library preparation from GEMs processing to the final 10x barcoded library. Picture courtesy of 10x Genomics.

1.2.2 Hi-C sequencing

Another sequencing method I employed during my thesis work is Hi-C, a high-throughput Chromosome Conformation Capture method based on Illumina sequencing[59]. With Hi-C it is possible to capture the three-dimensional spatial organization of chromatin inside the nucleus[59]. The chromatin is first crosslinked with formaldehyde to covalently form DNA complexes which are proximal in the 3D space. After a digestion step usually performed with a restriction enzyme, DNA fragment ends are bound to biotinylated nucleotides and fragments derived from the same complex are linked together to form ligation products (**Figure 2a**). A step of fragmentation (size fractionation) reduces the length of the DNA molecules to be suitable for short-read sequencing. Biotinylated fragments are purified using streptavidin magnetic beads and undergo a step of paired-end deep parallel sequencing, which includes the standard steps of adapter ligation and PCR amplification. Hi-C enables the identification of the origin of the sequences involved in the ligation product in the genome, but also their spatial organization. Each pair of Hi-C reads is chimeric and includes fragments that map to different genomic regions that are proximal in the 3D space[60]. The abundance of the coupled fragments is a function of the distance between any two given DNA fragments in the genome[60]. Since a chromosome occupies a particular territory inside the nucleus[61,62], spatially proximal DNA captured by Hi-C tends to originate from the same chromosome territory. If two sequences are close in the 3D space, they can be considered also close in the linear sequence. This principle can ultimately benefit the genome assembly process because it allows to link contigs or scaffolds[3] together and determine their order and orientation along the unknown chromosomes based on the interactions reported in the

[3]Contigs are gapless sequences. They can be joined together in longer sequences, called scaffolds, with gaps as separators.

reads, thereby generating more contiguous genomes. Hi-C data are also used to visualize genome spatial organization and validate the assembly of chromosomes through interactive contact heatmaps[59] (**Figure 2b**). A contact heatmap is a square symmetric matrix, in which rows and columns together represent the genome assembly divided into bins of fixed size[60,63]. Each entry in the heatmap represents pairs of interacting loci. Their position is determined by the location of either locus on the two axes, with different colors depending on their frequency value[63]. In a well-assembled genome, the diagonal shows a strong signal, with a high frequency of read pairs that are adjacent in the 3D genome sequence[60]. Intra-chromosomal interacting loci tend to create squares along the diagonal, representing entire chromosomes. If a Hi-C read contains interacting loci from different chromosomes, the entry will fall in a off-diagonal square corresponding to the two chromosomes (**Figure 2b**, green arrow), hinting at the presence of a missing join in the assembly, which could be merged in the same chromosome to increase contiguity[63].

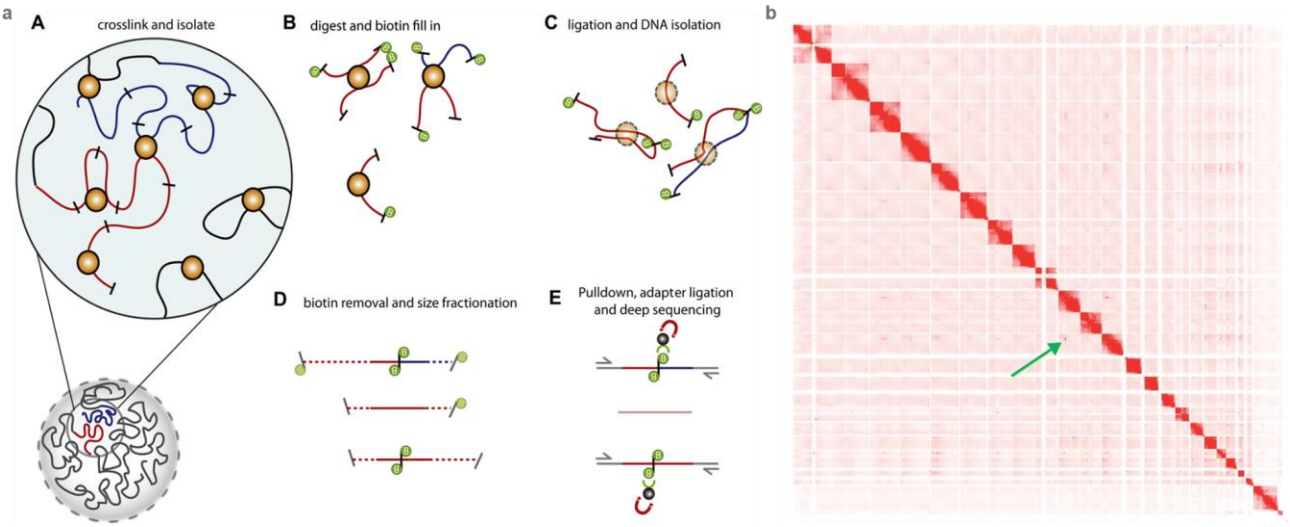


Figure 2. Hi-C sequencing technology. **a)** Overview of Hi-C library preparation. Figure is taken from Fig.1 of Belton et al., 2016[59]. **b)** Hi-C contact heatmap. Chromosomes are represented by pairs of interactive loci that fall into a square with a strong diagonal signal. The green arrow points at a pair of interacting loci that belong to different chromosomes but interact in the nucleus space. Figure taken from Figure 1a of Yardımcı et al., 2017[60].

1.3 Third-Generation Sequencing

To overcome the limitations imposed by Second-Generation Sequencing technologies, significant innovations in microfabrication led to new technologies, today often referred to as Third-Generation Sequencing (TGS). TGS technologies removed the amplification step, allowing the sequencing of single DNA molecules, a method called Single Molecule Real Time (SMRT) sequencing[64]. This approach generates long reads with a length >10 kbp on average[65] and is considered the best approach for *de novo* genome assembly[66]. Long reads can span entire repeats and can therefore resolve repetitive and complex regions[66]. TGS can also facilitate SV mapping[49,51,67]. Moreover, TGS can simultaneously detect epigenetic markers associated with the processed nucleotides[66]. A first issue of TGS approaches has been the cost per base pair compared to SGS, which, however, dropped

significantly in the last few years, allowing cost-effective genome assembly. The second issue is the higher error rate within long reads[68]. This limitation has been recently overcome with the latest PacBio SMRT technologies[69]. The two companies that provide TGS sequencing technologies are Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). The feasibility of Nanopore Sequencing was demonstrated in early 1990s, even before second-generation sequencing technologies became available, when researchers shown that single-stranded RNA or DNA could be driven across a lipid bilayer through large α-hemolysin ion channels (porins), which are normally employed in the passage of these kind of molecules[70–75] (reviewed in Branton et al., 2008[76]). In Nanopore Sequencing, which was not employed in this thesis work, tiny pores (nanopores) that in nature form gateways across membranes, are embedded into a synthetic membrane bathed in an electrophysiological solution (**Figure 3**). When a constant electric field is applied, an electric current can be observed in the system[77]. The double-stranded DNA is attached to an enzyme complex that approaches one of the nanopores. The single strand DNA is pulled through the pore one base at the time. As the DNA moves into the pore, the combination of nucleotides in the strand being processed creates a characteristic disruption in the electrical current[78]. This nanopore signal can be used to determine the order of the bases in the DNA strand and also epigenetic modifications in real time[79–81]. Nanopore sequencing provides the longest reads, with a record of a 2.3 Mbp single read[82]. Although being useful to assemble problematic regions thanks to their length, ONT reads are error prone, with a 5% to 40% error rate[83] and should be complemented with other technologies to increase the accuracy.

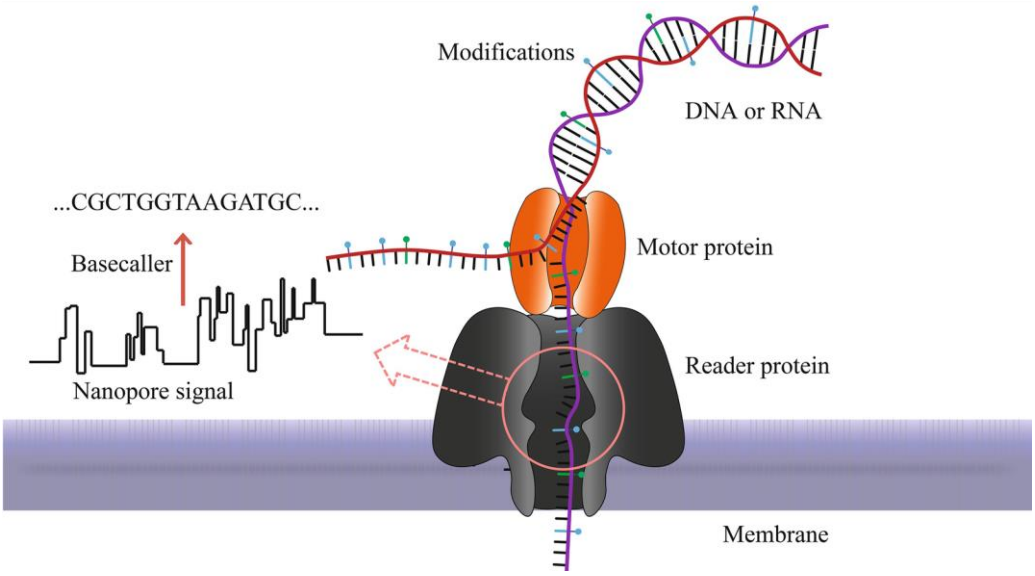


Figure 3. Nanopore sequencing. Figure taken from He et al., 2021[84].

In this thesis work, three different single molecule technologies were employed (see **Chapter 1**) and are described in the next paragraphs.

1.3.1 PacBio CLR long reads

In PacBio SMRT sequencing, double-stranded DNA molecules are ligated with specific hairpin adapters at both ends creating a circular template called SMRTbell[85]. In the sequencer, a single DNA polymerase and a single DNA molecule are present in the wells, called Zero-Mode Waveguides (ZMWs), of a SMRT cell[86,87](the reaction occurs for the 30-40% of the wells at best). The polymerase is immobilized at the bottom of the ZMW, where it binds to a primer complementary to the DNA adapter and starts the incorporation of fluorescently labeled nucleotides. The elongation reaction occurs in parallel for thousands of wells (**Figure 4a**). The wavelength emitted by each nucleotide incorporation is read in real-time to detect the correct base call. In Continuous Long Read (CLR) sequencing, the polymerase reads the entire circular template once (single-pass) to generate the longest possible read (**Figure 4b**). The length of these contiguous reads can be > 150kbp, being able to resolve difficult-to-assemble regions such as AT-rich or GC-rich regions, highly repetitive sequences, long homonucleotide stretches and palindromic sequences[88]. The reconstruction of such sequences is pivotal to build contiguous genomes, since the lack of coverage in these regions can leave gaps in the assembled sequence. Despite the advantages, the errors in CLR reads, mostly deletions or insertions[88], are around 10% and randomly distributed[13]. Errors are not systematic, and therefore can be resolved by consensus, i.e. by averaging the results from multiple reads (consensus, **Figure 4b**)[88].

1.3.2 PacBio CCS long reads

The newest PacBio sequencing strategy, released in 2019, implements SMRT PacBio sequencing with a new consensus approach that increases read accuracy (**Figure 4c**). Each double-stranded DNA molecule is circularized and sequenced multiple times, and a consensus is determined from the subreads generated from the same molecule[85,89] in a process called circular consensus sequencing (CCS)[69]. Resulting single molecule high-fidelity reads (HiFi) have an average length of 10-20 kbp and a base-call accuracy of over 99%, on par with short reads and Sanger sequencing[69]. HiFi reads establish a good compromise in terms of accuracy and length between short reads and long reads. HiFi reads made it possible to resolve difficult-to-assemble regions, generating highly-contiguous assemblies, and also to call small variants with a high confidence similar to short reads, while also detecting structural variation with higher confidence with respect to the CLR noisy long reads[69], without the need of additional consensus steps.

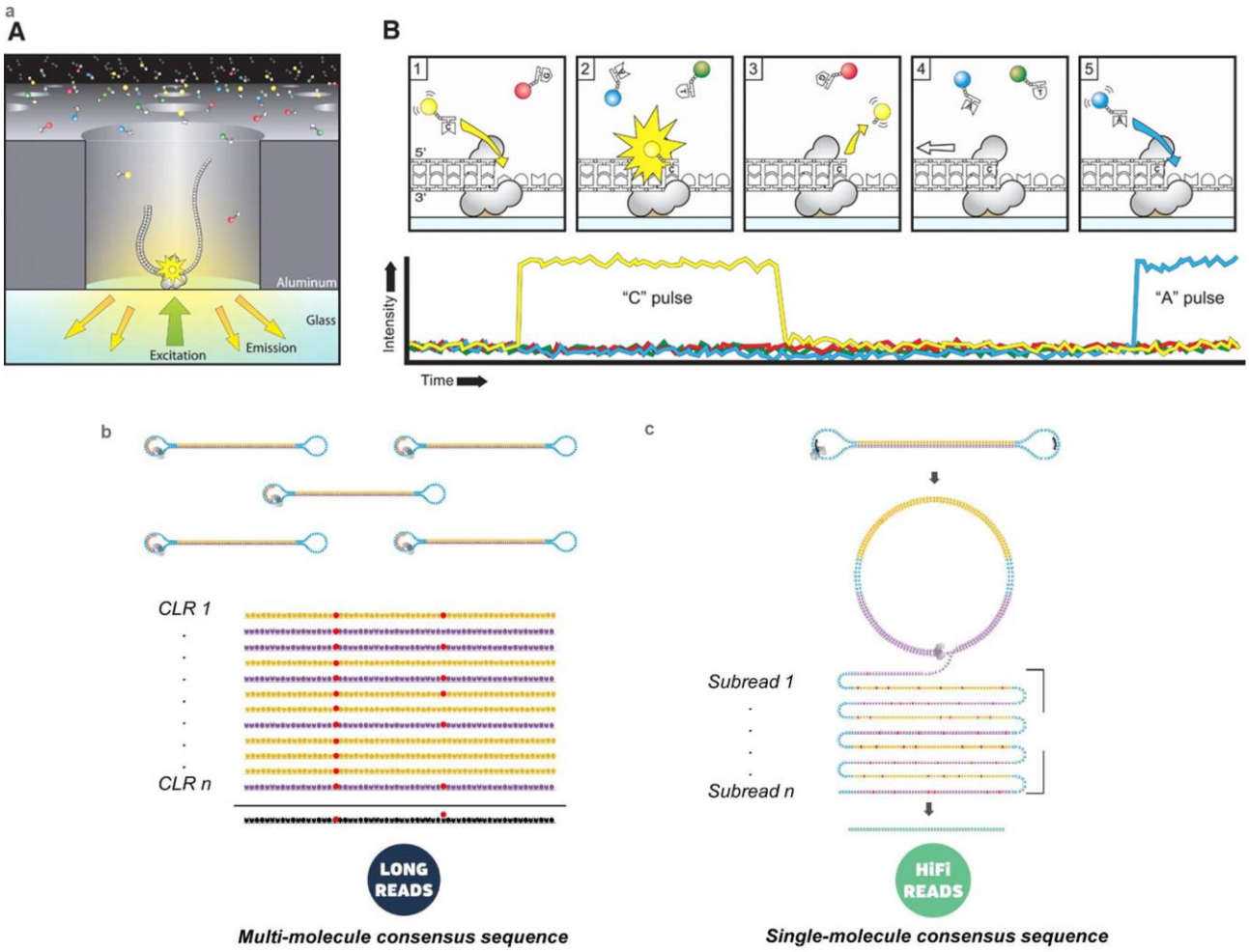


Figure 4. PacBio sequencing. **a)** SMRT technology employed by PacBio. The figure was taken from Eid et al., 2009[87]. **b)** PacBio CLR sequencing. Figure courtesy of Pacific Biosciences. **c)** PacBio CCS sequencing. Figure courtesy of Pacific Biosciences.

1.3.3 Bionano optical maps

Optical mapping is another single molecule technology commercialized by OpGen, Bionano Genomics, and NABsys (reviewed in Yuan et al., 2020[90]). In this approach, restriction enzymes are used to fluorescently label specific motifs in the DNA sequence. These sites create a unique fluorescent pattern on the DNA molecule that is read with image-capture softwares to create optical maps[90]. Bionano Genomics is the optical mapping technology used in this thesis work. Its technology relies on nanofluidic chips constituted of nanochannels in which the DNA molecules are kept uniformly elongated[91] (**Figure 5a**). The fluorescently labeled DNA molecules are driven into the nanochannels where they are automatically imaged with a high-resolution camera[91]. This method allows the generation of optical maps that include the physical locations of the fluorescent labels rather than base-level information[11,91,92] (**Figure 5b**). With respect to other optical mapping technologies, Bionano throughput and molecule length estimation was improved with the use of a more uniform linearization[91,93], and the preservation of molecule contiguity was guaranteed by using nicking enzymes that create single strand breaks (Nick, Label, Repair and Stain approach). With the

development of Direct Label and Strain (DLS) and the introduction of the non-nicking enzyme DLE-1, which is the approach employed in this thesis, Bionano optical mapping enhanced the map contiguity to the chromosome-level[90]. Optical maps have a molecule length (~255 kb) larger than both short and long reads[94] and can therefore span long repetitive or complex regions even further. Moreover, they report physical distances between selected DNA motifs and can therefore be used to correct errors in contigs or scaffolds joining and generate contiguous sequences during genome assembly[95], often allowing the generation of chromosome-level scaffolds[48].

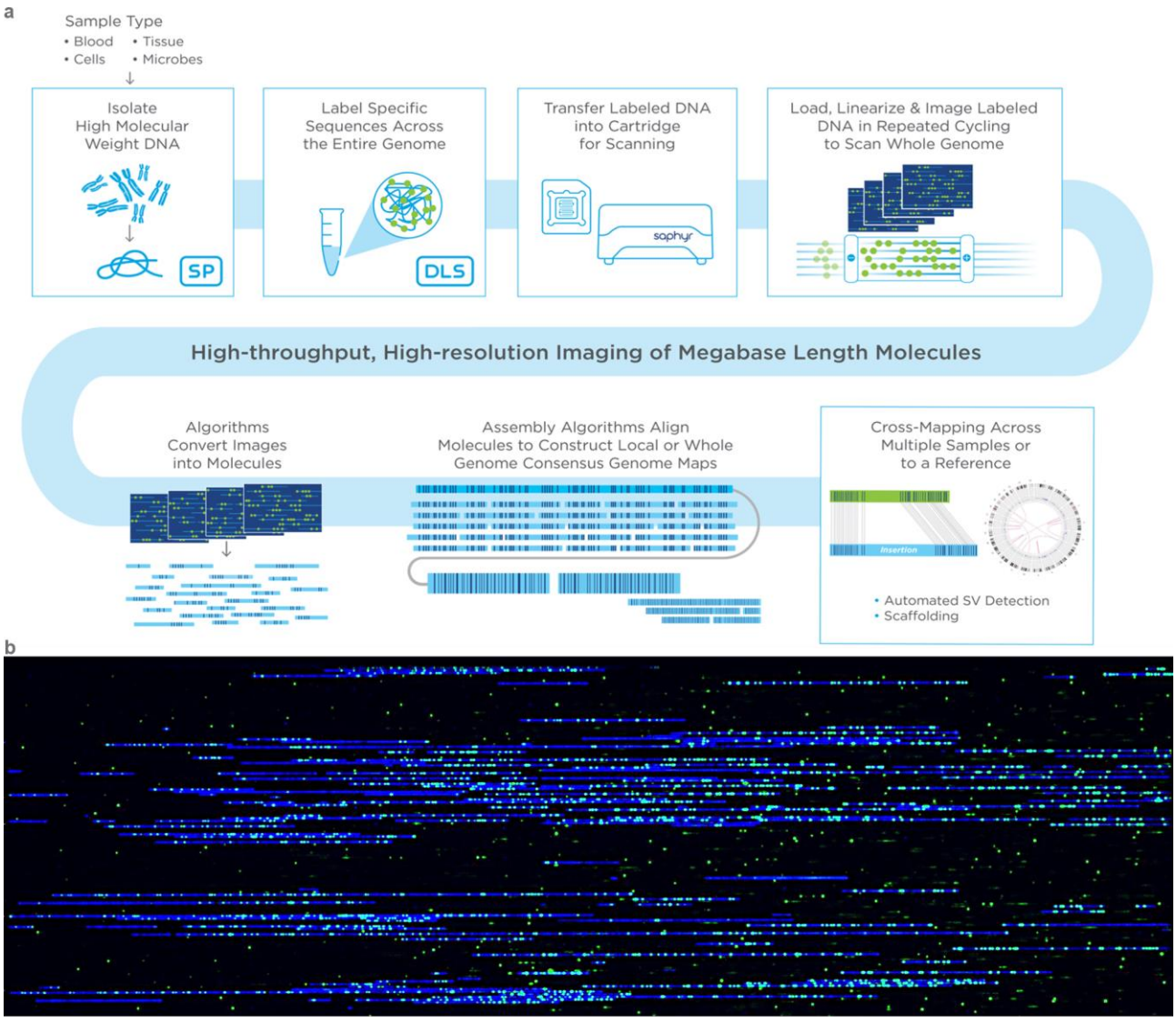


Figure 5. Bionano optical mapping. **a)** Process for the generation of Bionano optical maps. Picture taken from https://bionanogenomics.com. **b)** An example of a Bionano optical map. Blue segments are processed DNA fragments. Green dots are fluorescent-labeled restriction enzyme motifs. This graphic output can be visualized in real-time while generated from the machine. Picture courtesy of the Vertebrate Genomes Lab.

2. Generation of genome assemblies at scale

A reference genome can be regarded as the representative genomic sequence of a species. It serves as a standard set of sequences against which multiple users can refer their findings, promoting exchange of knowledge. Moreover, the availability of a reference genome reduces the costs of the analysis of other genomes of the same species, which can be based on an already existing sequence. The first eukaryotic reference genomes started to become available in the '90. Genomic data was initially generated with the expensive (1,500$ per Mbp today) and slow automated version of Sanger sequencing[96]. Therefore, early studies were conducted by large institutions which focused their attention only on a few model organisms[96]. A major effort was focused on the sequencing and assembly of the human genome, started in 1990 by the Human Genomes Project (HGP)[13]. The first eukaryotic genome to be sequenced was the yeast *Saccharomyces cerevisiae* in 1996[97], followed by the nematode *Caenorhabditis elegans*[98]. With the new millenium, the fruit fly *Drosophila melanogaster*[99], *Arabidopsis thaliana*[100], human (*Homo sapiens*)[101] and mouse (*Mus musculus*)[102] genomes were released. When second-generation sequencing methods (pyrosequencing, SOLiD, Ion Torrent and Solexa/Illumina) became available, the costs and time required to generate genome-scale data significantly decreased[96]. With SGS, a complete human genome (3.3 Gbp) could be generated in a single day, in contrast with the Sanger method, which took a decade to complete[103,104]. These technological advances allowed the generation of a large amount of cost-effective genome assemblies also by smaller laboratories, including the genomes of some non-model species[96].

2.1 Large consortia era: the beginning

Large multi-species sequencing projects were launched from 2009. For vertebrates, the Genome 10K (G10K) consortium, aimed to determine the genome of 10,000 vertebrate species, one for each genus[105]. Its main goal was to reconstruct the vertebrates' phylogeny, taking a deep look at the genomic changes that occurred during their evolutionary history. A unique feature of vertebrates is the presence of the neural crest, the fourth germ layer, which differentiates into a variety of nervous system cell types (glial cells, astrocytes, neurons), but also smooth muscle cells, tendons, adipocytes, melanocytes, craniofacial cartilage and skeletal elements[106,107]. Vertebrates' evolution is, therefore, spotted with great innovations like the presence of bones, cartilage and teeth, multi-chamber heart, sensory mechanisms and endocrine organs[108]. During their evolution, complex systems integrations combined with major anatomical and physiological changes, guided vertebrates colonization of a large variety of different ecological niches. As the human genome represented a fundamental resource to expand biomedical studies in the 21st century, a collection of multi-species genome assemblies is pivotal to boost biodiversity studies at all hierarchical levels[105]. The G10K genome collection aimed

at deciphering the evolution of vertebrates, enabling basewise studies in both coding and non-coding DNA[105], understanding how genomic changes shaped and generated their great biodiversity. Moreover, these genomic data will guide studies regarding recent vertebrate adaptations to environmental and climate changes, while also providing a powerful tool to boost conservation efforts of endangered species[109–111]. In the framework of the G10K project, the Avian Phylogenomic project and the Bird 10K (B10K)[112] project were established in 2014 and 2015 respectively, with the aim of sequencing all extant bird species genomes and collecting different biological data to promote avian genomics studies.

2.2 Avian genomics

2.2.1 Bird genomes

The avian class is the richest in species of tetrapod vertebrates, with 10,928 known species[113]. It includes domesticated species with agricultural interest, species important for neurogenetic and developmental studies and species studied for their role as pathogen vectors[114]. Birds are also widely used in conservation biology[115] and have a high cultural value for humans. The availability of enormous amounts of data on their biology and distribution, together with the vast catalog of known species, make birds suitable to answer challenging ecological and evolutionary questions[116,117]. The avian genome has many peculiarities that make them easier to analyze than other vertebrates. Among amniotes, birds have the smallest genome size (~1-1.4 Gb), which can facilitate large genomics studies[118]. This feature is related to the relatively low repetitive content[119] and generally shorter genes with respect to other tetrapods[120]. The ancestral avian genome also experienced the highest loss of sequence through deletion events[120]. Another feature of avian karyotype is its conservation, ranging from a haploid number of chromosomes between 76 and 80 in most species (http://www.genomesize.com). Some exceptions are found in the Accipitridae (2n = 50-68)[121,122], Falconiformes (2n = 40-52)[123], Psittaciformes (2n = 46-80)[124,125] and the stone curlew (*Burhinus oedicnemus*, 2n = 42)[126]. Major differences can be found in chromosome sizes within the avian genome. Three chromosome classes can usually be identified: macro-, intermediate- and microchromosomes, with differences in sizes also of an order of magnitude (from < 10Mb to 200Mb in chickens[110]). In the chicken (*G. gallus*), chromosomes smaller than 20 Mbp are considered microchromosomes[110]. Microchromosomes are present also in many reptilian lineages and they probably evolved 100-250 million years ago[127]. They are gene- and CpG islands-rich, with a higher recombination rate, while the repeat content is lower than that in larger chromosomes[110,128]. Microchromosome fusions have led to lower karyotype sizes in Accipitridae[121,122]. Another feature of avian genomes is that the heterogametic sex is the female, carrying two different sex chromosomes,

the Z and the W, which have low rates of recombination, while the male is homogametic (ZZ). The female heterogametic sex condition can also be found in butterflies, reptiles and amphibians, and it has evolved independently several times[129].

2.2.2 Early ornithological studies

Early ornithological genetic studies relied on the analysis of mitochondrial DNA (mtDNA), usually more abundant than nuclear DNA because present in more copies[130]. In the 1980s, the first studies involved the analysis of substitution patterns[131] and the use of the restriction fragment length polymorphism analysis (RFLP)[132] in mtDNA. These techniques were abandoned with the advent of PCR and DNA sequencing. For comparative studies, chromosome painting[4] and, later on, comparative mapping[5] were largely employed[133–136]. A strong evolutionary stasis of birds chromosomes was demonstrated by chromosome painting across multiple species[134–136]. In addition, comparative mapping, showed how the chicken genetic map is similar to that of the mammalian ancestor[133]. Large rearrangements that can affect the chromosome number and inter-chromosomal rearrangements are, therefore, considered rare events in avian genome evolution[129]. Despite these important achievements, early comparative maps with markers such as microsatellites[137] have been abandoned due to their low resolution. Indeed, in non-model species it was difficult to identify genomic markers for linkage studies even when a related model-species genome was available[129]. Later on, a new type of genetic marker derived from protein-coding genes became popular in ornithological studies, SNPs (Single Nucleotide Polymorphisms)[138,139]. SNPs can be integrated in linkage maps by genotyping in pedigrees and used for comparative studies[140]. Thanks to the advances in DNA sequencing, SNP-genotyping became the default method for linkage mapping in non-model bird species studies[129]. Linkage maps confirmed the evolutionary stability of bird chromosomes, with rare intrachromosomal rearrangement events, but with the occurrence of the intrachromosomal ones, in particular inversions and more complex SVs[141]. Despite the improvements achieved with genetic mapping, the resolution still remains suboptimal in detecting differences between species, with the risk of underestimation, and, therefore, the need for complete genomic sequences became compelling[129]. Whole-genome resequencing studies, which are widely used in population genomics in birds and other species, are also facilitated and refined by the usage of complete genomes. Resequencing projects rely on the mapping of sequencing reads against a high-quality and complete reference genome[13] to identify markers, mostly SNPs, to perform phylogeographic, conservation and adaptation studies. The advent of NGS technologies allowed the generation of complete genomes and

[4]Range of techniques that uses hybridization of fluorescently labeled DNA probes with cytological preparations to detect specific chromosome regions of rearrangements.
[5]Comparison of the location and order of homologous genes in different species.

the usage of sequencing approaches such as the common double-digest restriction-site associated DNA (ddRAD) sequencing, which provides sequence representation in restriction enzymes sites, but also whole-genome sequencing approaches, which enable the identification of more complete markers catalogs (see **Chapter 2**).

2.2.3 Mitochondrial genomes

Mitochondria are organelles which play crucial roles in eukaryotic cells. They supply energy in the form of ATP through aerobic respiration, generate macromolecules precursors (lipids, proteins, DNA and RNA) and reactive oxygen species (ROS)[142], also coordinating the cell response to oxidative stress, endoplasmic reticulum stress and DNA damage[143]. The mitochondrion evolved endosymbiotically from an prokaryotic progenitor[144] related to the extant alphaproteobacteria[145], which integrated into an archaea host cell[146]. Although many genes were transferred to the nucleus during endosymbiosis, the mitochondrial DNA constitutes a separate genomic entity to the nuclear genome. Indeed, the mitochondrion has an independent DNA replication, transcription and translation system, with a genetic code different from that of the nucleus[147]. The mitochondrial DNA (mtDNA) is double-stranded, and can exist in different conformations: linear, circular or branched[148]. In vertebrates, the mtDNA is a 14-20 kbp long circular molecule, usually harboring 37 genes that encode for 2 ribosomal RNAs (rRNAs), 13 proteins and 22 transfer RNAs (tRNAs)[149]. The vertebrate mitogenome has a single control region (CR) which includes the replication origin of one of the two strands and the origins of transcription for both strands[150]. The replication origin for the other strand is found between the *cox1* and *nad2* genes, within a cluster of tRNAs[150]. The CR is characterized by the presence of short and long repetitive regions and segmental duplications[151–153] that are harder to resolve[154,155]. To date, a large number of vertebrate species has its mtDNA sequenced[156], and new sequencing pipelines are being developed to resolve difficult-to-assemble regions[157]. Mitogenome-based research is routinely applied to a wide range of biological fields. Despite the present accessibility of the nuclear genome, the mitogenome, which can be considered as a single genetic marker, can still be valuable for several reasons[158]. Given its maternal inheritance, and therefore lack of recombination, the mitogenome can be used to complement nuclear studies and easily reconstruct maternal lineages[159]. Phylogeographic and phylogenetic studies benefit from the usage of mitochondrial genomes since they have a high mutation rate with respect to nuclear DNA[160], their gene set is strongly conserved across species, with few duplication and short intergenic regions[161], and are considered to have all the same gene tree[162] which is often close to the species tree[163]. Moreover, the phylogenetic resolution is improved from the large number of available sequenced taxa and individuals per species, which made the mtDNA trees often close to the nuclear topologies[158]. In birds, mtDNA is widely used to resolve taxonomy, phylogeny, ancient diversification and geographic

diversity (e.g. in Passeriformes[164] and diurnal raptors[165]), but also to analyze ancient species[166], parasites coevolution[167] and to guide conservation genetics[168].

2.2.4 Bird sequencing projects take off

The first sequencing studies in birds were focused on species of agricultural interest, with chicken (*Gallus gallus*) as the first sequenced bird genome in 2004[110], followed by turkey (*Meleagris gallopavo*) in 2010[169] and mallard duck (*Anas platyrhynchos*) in 2013[170]. Sequencing was also extended to non-model species for agriculture, but for other scientific disciplines[130]. The zebra finch (*Taeniopygia guttata*)[171] was sequenced in 2010 as the second bird sequenced after the chicken and was thought to boost avian genomics into the wild[172]. In 2012, the genomes of the pied and collared flycatcher (*Ficedula hypoleuca* and *F. albicollis*), and in 2013 the peregrine (*Falco peregrinus*), saker falcons (*F. cherrug*)[173] and the rock pigeon (*Columbia livia*) were released[174]. An important milestone in avian genomics was achieved in 2014, when Zhang and co-workers founded the Avian Phylogenomics Consortium (APC), with the goal of sequencing and analyzing a bird genome from each order and answering to several ecological and evolutionary questions, such as the relationships within Neoaves and their timing of radiation[175]. These questions can be answered with whole-genome analyses. However, a few annotated reference genomes were available at that time. In the initial APC effort, 45 additional avian species were sequenced using whole-genome shotgun sequencing for a total of 48 genomes, including chicken[110], turkey[169] and zebra finch[171], representing nearly all major clades of extant birds[120]. Genes were annotated using a homology-based method for all genomes, aided by transcriptomic data for some species[120]. A reference gene set with all genes from the chicken, zebra finch, and human was used to predict genes in all species and avoid biases related to the usage of different annotation methods[120]. Moreover, new bioinformatics and computational methods were developed for the study. The APC's analyses from this set of genomic data resulted in eight papers published in Science and 20 in other scientific journals[175].

One of the two flagship papers focused on a large-scale phylogenomic study: the reconstruction of the avian order phylogeny[176]. Earlier avian phylogenetic studies relied on morphological data[177,178], DNA-RNA hybridization[179], mitochondrial genomes[180,181], multiple gene fragments[182,183], larger sequences[184–186], and transposable elements insertions[187]. The usage of different analytical methods[188,189], data types[190,191], datasets[178,180,183,192,193], the incomplete lineage sorting (ILS) of the genes used[187,194], or a limited amount of data[195,196] in early studies resulted in contrasting trees. Thus, many questions and debates remained unresolved and the need to switch from phylogenetics to phylogenomics became necessary. The APC phylogenomic flagship paper[176] presented a well-resolved whole-genome tree, which confirmed some relationships proposed in earlier studies and

recovered groups that were not present[182,183,185,186], while also contradicting other studies[178–181]. In particular, it confirmed the separation of extant birds in three major groups: Neoaves, Paleognatae and Galloanseres, detecting also the first divergence of the Neoaves, which resulted in two sister clades: Passerea and Columbea. However, even with genome-scale analysis, some deep and short branches of Neoaves were not fully supported due to the presence of ILS during Neoavian basal radiation. Moreover, no single-gene tree was found identical to the species tree, suggesting that studies on single or multiple genes are not sufficient to resolve phylogenetic relationships. The source of this incongruence could be identified with more complete genomes, complete gene annotation between species and new methods[197]. Finally, this study corroborated the hypothesis of an avian rapid radiation during the Cretaceous-Paleogene boundary, which was associated with a high availability of ecological opportunities resulting from the asteroid impact and the subsequent habitat destruction and mass extinction[198–201].

The second APC flagship paper concentrated on the macroevolution of traits within the avian class[120]. The availability of complete genomes allowed to clarify several important features of the avian genome structure: its small size and gene condensation, the stasis of avian chromosomal evolution and the subsequent functional restraints. In particular, they found that 7.5% of the avian genome is under purifying selection (conserved). Zhang and co-workers[120] also conducted genome-wide association studies (GWAS) of convergent traits, focusing on vocal learner birds, while also investigating ecologically important candidate genes underlying avian traits. In particular, they analysed the evolution of flight through skeleton, pulmonary structure and feathers candidate genes, but also bird feeding modifications such as edentulism (absence of teeth) and dietary specializations with their related enzymes, as well as the genes involved in avian vision, reproduction and sex-related traits. Other papers included in the Avian Phylogenomic wave of publications, investigated in details tooth loss[203], blood and platelet cell genes[204,205] and vocal learning[206–208].

This collection of papers is a perfect example of the huge amount of high-quality genome data needed to reconstruct the phylogeny of a clade that radiated so rapidly, and to analyze its micro- and macroevolution through comparative genomics[120,176]. However, these studies also indicated that more genomes are necessary to completely resolve the avian tree[176]. Moreover, when a deeper detail at the gene level was necessary, those assemblies failed to completely fulfill this purpose. Indeed, misassemblies and gene incompleteness were found in the APC assemblies[209]. In 2015, the B10K consortium[112] stemmed from the APC efforts, with the aim of sequencing all the extant bird species. In 2016, a study on avian genome evolution established new methods for the study of synteny and rearrangements between species[210]. They used the genomes generated in Zhang et al 2014[120], pointing at the importance of having genomes as complete as possible, with entire chromosomes assembled[210].

In 2017, PacBio single molecule, high-throughput long-reads sequencing methods were employed to improve short and intermediate-reads assemblies and evaluate the standards for G10K and B10K projects[209]. In particular, Korlach and coworkers[209] focused on two bird genomes: the zebra finch, already assembled with Illumina short reads, and Anna's hummingbird (*Calypte anna*), previously assembled with Sanger intermediate reads. The two long-read assemblies resulting from the study[209] have an increased genome contiguity, with a decrease in contigs number and increase in contig N50[6]. Furthermore, the new assembler (FALCON and FALCON-unzip[211]) allowed to phase diploid genomes, separating the two haplotypes and resolving allelic differences and assembly errors between divergent haplotypes. The new methods also filled assembly gaps and resolved sequence errors surrounding them, demonstrating that long-reads assemblies have a higher base calling accuracy. Regarding gene completeness, less fragmentation was found with respect to the older assemblies, improving gene prediction and reducing missing genes[209].

In the framework of the B10K project, a new method for comparative genomics was released in 2020 called Progressive Cactus[212]. Progressive Cactus is a reference-free multiple genome alignment tool that avoids the reference bias present in common alignment programs. Cactus relies on an input tree to progressively align the genomes. This approach divides the alignment process into sub-alignments which are solved independently. Cactus also allows reconstructing ancestral genomes at each node[212]. This new method was applied in a recent study by the B10K consortium[213] using 363 bird genomes sequenced with short reads during phase II of the project. The elevated number of species increased the statistical power to detect evolutionary constraints, revealing 13.2% of the chicken genome as conserved[213]. However, comparative studies at the gene level will benefit not only from the number of species used, but also from the completeness and contiguity of the genomes.

2.3 The "Big Data" era

We live in the era of big data. For genomics, this means hundreds of genomes being sequenced worldwide every year. By October 2017, in the National Center for Biotechnology Information (NCBI) database, only 2,534 eukaryotic species genomes were available and only 25 reached the goals in terms of contiguity set by the G10K consortium[214]. This represents only a tiny fraction of the eukaryotic Tree of Life. Many genomic features can be discovered and described only by comparative analysis between close sister species[215], while correct and wide phylogenetic trees will provide fundamental support to understand phenotypic diversification[216]. Other important biological

[6]N50 is measure of assembly contiguity, indicating that 50% of the assembly is represented by contigs/scaffolds with length equal or higher than the N50 value. N50 is calculated in the context of the assembly size rather than the genome size, therefore it is often replaced by the NG50, which is conceptually the same as N50, but it relies on genome length, allowing more meaningful comparisons between different assemblies.

questions that we will be able to answer with a large collection of sequenced species include the origin of eukaryotic cells, genomic changes during symbiosis, chromosome evolution, eukaryotic gene regulation, diversity of sexual systems, speciation, complex traits development, ecosystem function, stasis and change, and species conservation[216]. From 2017, the need to generate chromosome-level assemblies, as learned from the APC lesson, pushed the establishment of new sequencing projects such as the Vertebrate Genomes Project (VGP)[48], the Bat1K project (B1K)[217], the Earth Biogenome Project (EBP)[214], the Darwin Tree of Life Project (DToL)[215], the European Reference Genome Atlas (ERGA)[218] and the Telomere-to-Telomere consortium (T2T)[219,220]. For this thesis work I was involved in three of these projects, which I describe in the next paragraphs.

2.3.1 The Vertebrate Genomes Project (VGP)

The Vertebrate Genomes Project was established in 2017 with the aim to sequence a near error-free, chromosome-level reference genome for all the ~70.000 living vertebrate species (https://vertebrategenomesproject.org/)[48]. The genomes will be publicly available in the public archives such as NCBI (Accession: PRJNA489243). They are also currently shared through a dedicated VGP repository (https://vgp.github.io/genomeark/). The project is divided into four different phases according to taxonomic classification: Orders (Phase I), Families (Phase II), Genera (Phase III) and species (Phase IV). The VGP is currently in phase I, which involves the sequencing of 263 vertebrate Orders and 4 invertebrate outgroups. As of February 17 2022, 116 genomes had been completed, 62 were in progress and 44 had funding but not samples, yet, leaving only the 10% of the Orders without a reference genome. Several biological questions that can be answered with the availability of all vertebrate species: the reconstruction of a genome-scale tree for vertebrates, comparative genomics studies on both specialized and convergent traits like vocal learning, investigation of vertebrate chromosome evolution, reconstruction of the ancestral genome of all vertebrates and single clades, and conservation biology planning. To date, the VGP boasts a series of publications regarding new developed methods for genome assembly[48,221], genome curation[222] and mitochondrial genome assembly[157], but also reference genomes and biological discoveries papers[125,223–227]. These papers were published in a special issue of Nature in 2020 and other high-ranked journals. I will now give a brief outline of the content of each of these papers. The VGP flagship paper[48], which presented a dedicated pipeline for the chromosome-level assembly of vertebrate species and a pipeline to generate completely phased assemblies, will be the subject of **Chapter 1**. An important step in the VGP assembly pipeline is the presence of a final genome curation step, before gene annotation[48]. This step is important to correct (some) of the errors left by the automated assembly algorithms[222]. Indeed, repetitive or highly heterozygous regions can be difficult to resolve relying only on the assembly of contigs from raw reads and the subsequent scaffolding

process. Looking at the correctness and completeness of an assembly, and not only at its contiguity, false duplications, misjoins, missed joints, and collapsed regions, can be present in the final assembly generated from automated pipelines. The gEVAL[228] browser was generated to visualize such errors using all the raw data employed in the assembly process. After the visualization, manual work is performed by experienced curators with dedicated pipelines to generate the final improved assembly[222]. Another VGP methodological paper presented a new tool, FALCON-phase, to help phasing the two haplotypes of a diploid genome in an innovative way with respect to previous efforts[221]. It employs Hi-C data from the sequenced individual to generate parental haplotig blocks (i.e. blocks of sequences that belong to the same maternal or paternal homologous chromosome). The MitoVGP pipeline paper[157] presented a new pipeline to assemble mitochondrial genomes from PacBio and ONT long reads, and 10x Genomics Linked-Reads. The mitogenomes assembled with this pipeline are more complete than those assembled with just short reads, being able to represent mitochondrial repetitive regions and duplications[157]. The VGP wave of publications in 2020 also includes reference genome papers with biological discoveries. The VGP gives priority to endangered or near to extinction species. The genome of the vaquita (*Phocoena sinus*), a critically endangered small cetacean, was generated to deepen the knowledge on its demographic and evolutionary history[224]. They confirmed that the vaquita's small population size was not the result of a recent bottleneck or inbreeding. Given that its small population size persisted for 300,000 years, the low-heterozygous genome had the opportunity to purge deleterious alleles and keep the population at equilibrium, maintaining the necessary diversity. This implies that potentially population recovery will not be hampered by its genetics[224]. Another endangered species targeted by the VGP is the kakapo (kākāpō, *Strigops habroptila*), a flightless parrot endemic to New Zealand with less than 200 extant individuals[125] (https://www.doc.govt.nz/our-work/kakapo-recovery/). Similar to the vaquita, small kakapo populations could survive when isolated for hundreds of generations[125]. The current population, derived from individuals that were isolated on a small NZ island ~10,000 years ago, shows less deleterious mutations than the extinct mainland individuals, a condition probably due to genetic drift and mutation purging through inbreeding and purifying selection[125]. The study underlined the critical importance of identifying deleterious mutations in the kakapo population and to evaluate the effect of mixing the birds of mainland genetic heritage with the island ones. This could help to genetically rescue the highly inbred population, but can also introduce deleterious mutations from the higher-mutational-load mainland lineage which could be easily fixed in this extremely small population. Both studies shed light on the evolutionary history of endangered species and highlighted the importance of genomic data, helping to determine the strategies needed for their preservation[125,224]. Another VGP-related study focussed on the genome of the marmoset (*Callithrix jacchus*), shedding light on its structure and evolution, in particular regarding the Y chromosome, and

established some precautions to be taken when using the primate as a model for biomedical studies[226]. The genomes of the only two extant monotremes, the platypus (*Ornithorhynchus anatinus*) and the echidna (*Tachyglossus aculeatus*), instead provided insights into their peculiar characteristics and mammalian genome evolution[227]. The chromosome-level genomes permitted to better understand the evolution of their multi-X and multi-Y sex chromosomes, and the evolution of traits different to therian mammals, which shaped the ecological adaptation of monotremes[227]. Another paper presented the chromosome-level genomes of six bat species, providing insights into their adaptations like flight, echolocation, immunity and longevity[223]. In particular, this study, and in general the Bat1K efforts[217], are willing to unveil the molecular mechanisms at the basis of bats' exceptional immunity system and longevity, knowledge that can be applied also to human health. For example, in the future, bats' tolerance to coronavirus may be relevant to increase human survivability in the current COVID-19 pandemic[223]. Lastly, a study on oxytocin, vasotocin and their receptors, proposed a universal nomenclature for the latter, in order to avoid confusions related to earlier incomplete genomes and gene sets[225]. Through this first wave of publications, the VGP established the importance of having chromosome-level assemblies in all biological fields, providing complete genomic sequences and gene sets that can be used to address the most challenging biological questions.

2.3.2 The Darwin Tree of Life (DToL)

The Darwin Tree of Life[215] is one of the key activities of the Tree of Life program of the Wellcome Sanger Institute (https://www.sanger.ac.uk/). It is an initiative established in the contest of the Earth Biogenome Project, whose final goal is to sequence, catalog and characterize the genomes of all of Earth's eukaryotes[214]. The DToL aims to sequence at the chromosome-level all the about 70,000 eukaryotic species of England and Ireland (https://www.darwintreeoflife.org/), providing key contributions also to other sequencing projects, such as the Vertebrate Genomes Project[48]. The habitats of England and Ireland are impacted by the influence of different drifts, streams and currents[215]. The biodiversity of these islands, which mostly resulted from recolonization after the first glacial maximum[229,230], gives the opportunity to study the response of species to climate changes, but also the effect of species invasion and human presence[215]. Indeed, the territory and its biota have been affected by thousands of years of anthropogenic changes, including deforestation and pollution[215].

2.3.3 The European Reference Genome Atlas (ERGA)

ERGA is a pan-European initiative initially discussed in December 2020 to generate a reference genome for all the ~200,000 eukaryotic European species, including endemic, keystone and endangered species, but also species with an economical and ecosystemic importance. ERGA goals fall within the European Commission priority to conserve and restore biodiversity (European

Commission Biodiversity Strategy) and the Horizon Europe program (see www.erga-biodiversity.eu). ERGA was established as the European node of the Earth Biogenome Project, involving cross-country infrastructure and coordination from nearly 50 EU countries and other partners. Standards for genome quality and contiguity were set similar to those of EBP[214] and DToL[215] to generate near complete and near error free reference genomes, enabling a solid and comparable foundation for all conservation, monitoring and restoration projects across Europe.

2.4 Emerging applications for complete reference genomes

2.4.1 Pangenomics

The generation of the human reference genome by the Human Genome Project has tremendously advanced genomic research. We are now at the 38th update of the human reference genome (GRCh38), which is continually being improved through gap filling and repetitive regions resolution with the implementation of long reads and the latest HiFi reads[69]. In 2021, The Telomere-to-Telomere Consortium (T2T)[7] announced that they filled all the gaps in each human chromosome, presenting the most complete version of the human genome[220]. Population-level studies will surely benefit from this extremely complete resource, allowing the construction of the first comprehensive catalog of genomic variants[231]. However, it has become clearer and clearer that a reference genome constructed from a single individual is not enough to represent the variability that characterizes an entire population[232–234]. Indeed, a reference usually comprises genomic sequences, and therefore allelic diversity, specific to the specimen[235]. This was noticed with the first human personal genomes sequenced in 2007[103,104]. These observations led to the establishment of sequencing projects as the HapMap[236,237] and the 1000 Genomes Projects[238]. Using the 1000 Genomes Projects individuals, it was found that around 2 millions of reference alleles are "minor alleles", with population frequencies of less than 0.5[235]. Looking at the allele distribution of the reference, it is very similar to that of the other sequenced personal genomes, meaning that the current human reference can be considered a well-defined and well-assembled haploid personal genome[235]. In the last few years, additional human genomes from different ethnicities were released (reviewed in[239]), demonstrating that up to 10% of their sequences were missing from the current reference. Therefore, conducting research studies using a traditional linear-reference genome can introduce a significant reference bias[235]. The availability of high-throughput sequencing techniques and high-quality reference genomes helped resequencing projects, which usually rely on the sequencing of multiple samples followed by the mapping of the data against a single reference[240]. However, despite this being a standard approach, reads that are not identical to the reference can fail to map, affecting the variant calling process[233,241]. Indeed, when the

[7] A Telomere-to-Telomere genome has completely assembled chromosomes end to end, without gaps.

reads have non-reference alleles, variant callers can fail to detect some variants or overestimate them[233,241]. For example, in clinical studies, when the reference comprises rare alleles, pathogenic variants can be ignored and considered as benign[235]. These issues, usually referred to as "reference bias", can affect SNPs detection[242,243], but also structural variation assessment using both short and long reads[49,244], leading to the characterization of the reference sequence itself rather than the entire population under study. Therefore, the development of a new method to collect all the variability that characterizes a species became urgent.

Pangenomics is a new research area which aims to replace the linear reference genomes with a more complex genomic structure. This is meant to overcome common issues that metagenomics, comparative genomics and population genomics face when trying to represent and analyze the entire variability of a species[245]. The concept of pangenome was first proposed in 2005 for bacterial species as a collection of core genes that are shared between all the individuals, dispensable genes shared between some individuals (accessory genes) and those that are unique to one strain (unique genes)[246]. For eukaryotes, since ~50% of the genome is composed of intergenic sequences (non-coding), and genes can harbor long introns[247], the definition was expanded not only to genes, but to the whole genomic sequence. The first eukaryotic pangenomes were generated for food crop species (reviewed in Sherman and Salzberg 2020[239]), in order to find the relationships between the absence or presence of genes and important phenotypes in particular varieties. Several projects were launched to generate catalogs of genomic variation in humans and capture novel sequences in different populations, but without the final aim of generating a pangenome[239]. In 2019, the Human Pangenome Project was established as a worldwide effort to generate telomere-to-telomere assemblies for hundreds of human individuals and generate a human pangenome[231]. Only few pangenomes were generated for other animal species with an economical relevance, such as Mediterranean mussel (*Mytilus galloprovincialis*)[248], the domestic pig (*Sus scrofa domestica*)[249], the Simmental cattle (*Bos taurus taurus*)[250] and the chicken (*Gallus gallus*)[251,252].

Computational methods that ease storage and retrieval of genetic information in a pangenome have only recently been developed. The current approach is to represent the reference and the variants as a graph, with paths and nodes as main elements. Paths represent the sequences included in the pangenome which underlie the graph data structure. Paths walk along multiple nodes, which are specific sequences of bases that can be identical between paths, or different, indicating the presence of variants. Available pangenomics bioinformatics tools are SevenBridges[253], PaSGAL[254], GraphAligner[255], Cactus Pangenome Pipeline, pggb and vg[233,256]. These methods allow the construction of pangenome graphs from a collection of multiple reference genomes, haplotype

sequences, variants linked to the references, and raw reads alignments[245]. Other types of information can also be embedded in the graph, such as gene annotation, haplotype information, epigenetic properties and species taxonomy[245]. A pangenome graph can be visualized with tools such as odgi[257] or SequenceTubeMap[258], being also able to graphically represent variants. The pangenome can then be analyzed, to search for example for core and dispensable genes, or used as a richer reference genome for variant calling, allowing also the incorporation of new data. The use of a pangenome can improve the read mapping and variant calling, in particular in regions enriched in variants[245]. By considering already detected variants, read alignment can be performed in a variant-aware way[259]. Indeed, the subsequent variant calling can identify both known and novel variants[245]. When a read maps to embedded variants in the graph, these are considered variants present in the donor that have already been identified. On the other hand, when not already present in the graph, they can be considered as novel variants like in a classic variant calling approach. Pangenome graphs also allow the representation and detection of complex variability such as nested variants (e.g. SNPs included in a large SV)[239]. Resequencing projects based on a pangenome reference should therefore overcome the limitation imposed by the reference bias[240].

2.4.2 Conservation genomics

Our planet is currently experiencing the sixth mass extinction[260]. Habitats are being destroyed to meet the needs of the constantly growing human population, and climate change is markedly increasing due to the massive consumption of fossil energy sources. Conservation genetics is rapidly shifting to genomic approaches. Indeed, genome-scale data can provide a wider range of markers that can be used to precisely estimate endangered species genetic diversity, population structure and demographic history (reviewed in[261]). However, conservation genomics has been hindered by the lack of available genomes for most of the species[261]. Recent development of cost-effective ways to sequence and assemble complete genomes, together with the establishment of sequencing consortiums, paved the way to a new era of conservation genetics (see **Chapter 6**).

3. Model species

Among the bird species for which I assembled a chromosome-level reference genome during my PhD, I will focus on three species: the barn swallow, the European nightjar and the lesser kestrel.

3.1 The barn swallow

2.1.1 Biology

The barn swallow (*Hirundo rustica*, **Figure 6**) is a synanthropic and insectivorous passerine bird that belongs to the Hirundinidae family, which comprises swallows and martins[262]. Its body is small and aerodynamic, with a length between 17 and 19 cm and a weight around 20 g, while the wingspan is up to 35 cm long[262,263]. Barn swallows have contrasting metallic blue upperparts and white to rufous ventral parts, while forehead and throat are red[264]. Sexual dimorphism is slightly marked, with the most noticeable difference being the tail length. Indeed, adult males have longer outermost tail feathers than females (on average 104 mm vs. 98 mm)[265], while this difference is not present in chicks and juveniles before the first complete winter molt. Clutches include 1-7 eggs which are incubated by females for 14 days[266], and the same pair can have multiple broods in the same reproductive season. The natal dispersal is high, with only 5% of returns to the natal site for the breeding (mostly males)[262,263,267,268]. Conversely, adults of both sexes have a high breeding philopatry, with the tendency to return to the same breeding site[263]. The barn swallow is a polytypic species, with 8 recognized subspecies: *H. r. rustica* (Europe, North Africa and Western Asia), *H. r. savignii* (Egypt), *H. r. transitiva* (Israel, Lebanon, Jordan and Syria), *H. r. tytleri* (southern-central Siberia, Mongolia), *H. r. gutturalis* (central-eastern China, Japan), *H. r. erythrogaster* (North America) and the debated asian subspecies *H. r. saturata* and *H. r. mandshurica*. The majority of the subspecies are migratory, with the exception of the resident egyptian (*H. r. savignii*) and Middle East (*H. r. transitiva*) subspecies[262,264,269]. Barn swallows breed in Eurasia and North America, while wintering

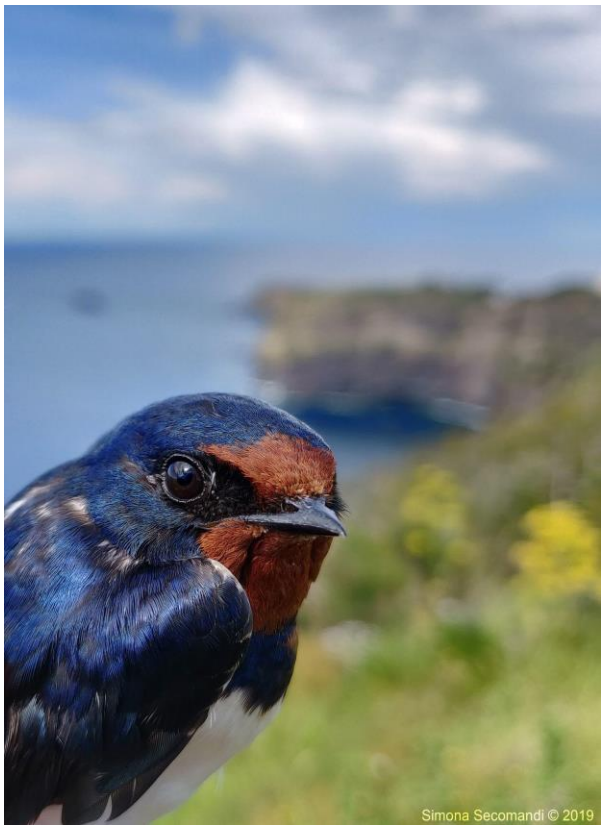


Figure 6. A barn swallow individual sampled in Ventotene, Latina (Italy) in 2019 during my first PhD year.

territories are found in Africa south of the Sahara Desert, northern Australia and South America[262,270]. In particular, the European subspecies winters from Sahel to South Africa[262,271].

2.1.2 Association with humans

Barn swallows have a wide holarctic distribution, being the most widespread species in the swallow family. Their great expansion can be linked to its ability to exploit anthropogenic environments, in particular agricultural structures such as barns, stables, and cattlesheds[272]. The barn swallow can be considered a synanthropic species, which have adapted to exploit human environments without the need of an obligate dependency on anthropogenic resources[273,274]. Advantages of its synanthropic behavior include the availability of nesting sites, but also that of insects surrounding lifestocks[275]. Its association with human settlements seems to have lasted for millennia[272]. A recent study outlined this hypothesis, linking a 7,000 years ago bottleneck event in the ancestral population with the spread of early human settlement[276]. A subsequent founder effect may have coincided with the establishment of barn swallows' association with humans, with few founder individuals that colonized anthropogenic structures before subspecies divergence, later expanding to occupy their Holarctic range[276]. The barn swallow species complex is not endangered, but it experienced a decline at the population level, mostly related to agriculture conditions, such as variations in livestock farming[277] and the decrease of flying insects due to the use of pesticides[278].

Synanthropy can be considered as a less strict association with humans compared to domestication. Domesticated species are actively selected by humans, which enhance their survival and breeding, and, in advanced stages, control selective pressures directly[274]. A primary selective pressure during early domestication stages targets behavioral mechanisms such as tameness, the attenuation of aggressiveness against human caretakers, which is a pivotal prerequisite for a successful species domestication[279,280]. Its onset is the result of the reduction of the adrenal glands, which control fear and stress response. This modification, together with all the other domestication traits such as floppy ears, smaller teeth, depigmentation and curly tails, accounts to Darwin's "domestication syndrome"[281]. One hypothesis on the molecular mechanisms underlying such traits, suggests a role of the neural crest cells (NCC), vertebrate-specific multipotent stem cells that arise from the embryos dorsal neural tube[281]. In particular, it was suggested that the selection for tameness causes a reduction of NCC-derived tissues with a behavioral relevance, also generating the other "domestication syndrome" phenotypic traits[281]. The bases of this reduction are genetic, involving many alleles across many genes influencing the development, migration and interaction of NCC under the "domestication syndrome"[281]. A central role in the onset of domestication traits was found in the glutamatergic system (reviewed in O'Rourke and Boeckx, 2020[282]). Glutamate is the main neurotransmitter in the vertebrate

central nervous system, with an excitatory function in stress circuits in the hypothalamus[282]. It is involved in learning, memory, plasticity and the regulation of the hypothalamic–pituitary–adrenal (HPA) axis[283]. Signatures of selection in domesticated species affect genes involved in the regulation of glutamate signaling[284]. Changes in genes for glutamate receptors were found implied in the reduction of the stress response, and therefore of aggressiveness, under domestication[282]. Some of these glutamate receptor genes were also found to be targeted by *foxp2* in the brain of songbirds[285]. *Foxp2*, together with *foxp1*, are well-studied paralogous transcription factor genes involved in neurodevelopment and vocal learning circuits[286–288]. Changes in the glutamatergic system under domestication have been found to be involved in vocal learning mechanisms, with more variable vocalizations in domesticated species than in the wild ones (see O'Rourke et al., 2021[284]). A synanthropic vocal-learner species with a long-lasting association with humans, such as the barn swallow, can be considered a good model for the study of the genetic basis of synantropism and vocal learning.

2.2.2 Genetic studies

The barn swallow is one of the most studied birds in the world with an enormous collection of studies on its behavior, ecology and evolution (e.g.[289–294]). However, it has been only in recent years that its genetics started to be explored, mostly because of the lack of reference resources. Earlier studies concentrated on the genetic control of phenotypic traits, in particular related to migration and the circadian clock. The timing of annual life-history events (i.e phenology) is synchronized with seasonal changes in ecological conditions that occur during the year, especially at medium and high-latitudes, where seasonal fluctuations are more accentuated[295]. Seasonal activities are controlled by an endogenous circannual and circadian rhythmicity[295], which is modulated by the endogenous clock, a biochemical oscillator that sets the timing of behavioral and physiological processes related to breeding, molt and migration[295–297]. This endogenous clock has a strong genetic component[297] and is influenced by exogenous cues[298] of which the most important is photoperiod (i.e. circannual variation in the relative duration of day and night), that allows birds to time their behavior according to seasonal ecological changes[299]. Migratory birds show a strong phenotypic variation of migration timing among individuals and populations. Experimental studies that have focused on quantitative genetics and heritability of timing of migration have shown that this trait has a large additive genetic component[300]. It has been suggested that phenological variation in migration timing may depend on polymorphisms at genes involved in the signaling cascade of the endogenous clock[297]. However, little efforts have so far been devoted to identifying the genes involved in shaping phenotypic variability in natural populations of migratory species[299] and only few genes have been putatively associated with this variation (e.g. *Clock*, *Adcyap1*, *Npas2* and *Creb1*[271,301–307]). The availability of a reference genome

for the barn swallow and an accurate gene annotation presented here will definitely help to get insight into the genetic basis of this complex behavior.

More recent genomic studies focused mostly on barn swallow populations' genetics, sex related traits, hybrid zones, and differences between subspecies. In 2016, a first draft genome for the North American subspecies (*H. r. erythrogaster*) was generated to investigate the contribution of selection and geographic isolation to genome-wide differentiation in 8 closely related barn swallows populations[308]. Using ddRAD sequencing data and the draft genome, the authors identified around 9,493 SNPs. The authors concluded that morphological adaptation, related to migration and sexual signalling behaviours, shaped population-level differentiation. In the same year, eight microsatellite loci developed for the barn swallow[309,310], stretches of the *nd2* mitochondrial gene and more than 20,000 ddRAD markers were used to assess selection pressures associated with divergent migratory phenotypes in *H. r. rustica* sympatric populations across the migratory divide in Central Europe[311]. The authors found no population structure across the different populations, suggesting a large panmictic population subjected to gene flow, with only one genetic cluster explaining the genotypic data. Wilkins and collaborators in 2016 demonstrated a significant phenotypic differentiation among sexual signalling traits in two subspecies, *Hirundo r. rustica* and *H. r. transitiva*, which show low genetic differentiation probably due to gene flow between populations[312]. They used SNPs generated using a genotyping-by-sequencing (GBS) approach and the barn swallow draft genome[308]. Very recently, the genetic effects of migratory divides were further evaluated by the same group[313,314]. In 2017, another ddRAD-based study concentrated on hybrid zones[315]. Using >23,000 SNPs from *H. r. rustica*, *tytleri* and *gutturalis* among a transet in Siberia, the authors measured gene flow between the three subspecies and found that the degree of divergence in the ventral coloration and wing tail, which are targeted by natural and sexual selection and differ in the subspecies, are associated with the extent of hybridization in secondary contact zones. The three subspecies differentiation was also linked to genomic regions associated with throat brightness and wing length. A 2018 study, which was already discussed in the last paragraph, outlined the association between barn swallow population expansion in the Holarctic and the spread of human settlements[316]. They performed whole-genome sequencing on eight *H. r. savignii* and eight *H. r. erythrogaster* individuals, and genotyping-by-sequencing on barn swallow individuals from all across the northern hemisphere. A recent paper published in 2021, shedded insights into sex-related genetic diversity, which is shaped by selection, gene flow and demography[317]. They also outlined the importance of the Z-chromosome in speciation events. To this end, they generated Illumina data for six barn swallow subspecies and aligned them to a new barn swallow reference genome[318] for variant calling. This new reference was generated in 2019 by my research group at the University of Milan with a combination of PacBio CLR long reads and Bionano

Optical maps which enabled the assembly of a contiguous scaffold-level reference for the European barn swallow (*H. r. rustica* subspecies)[318], whose quality metrics far exceed those of the assembly generated with Illumina paired-end short reads in 2016[308]. Indeed, the latter was far from being contiguous, with an N50 of 0,39 Mbp[308], while the new N50 was 26 Mbp[318]. In 2020, a new reference genome for the barn swallow was released by the B10K[319] and included in a huge comparative genomics study[320]. Its N50 was lower than that of the 2019 assembly (0,68 Mbp), once again not providing the necessary contiguity for accurate genomic studies on the species. Despite these recent efforts, the chromosome-level was not reached because of the sequencing technologies and assembly pipelines used. As part of this thesis work, I generated a new chromosome-level reference genome for the barn swallow (see **Chapter 2**) using the Vertebrate Genomes Project assembly pipeline (see **Chapter 1**).

Phylogenetic studies on barn swallows were also performed by using mtDNA[272,321–323]. These studies highlighted that the barn swallow complex is monophyletic and that its great Holarctic distribution was reached after the separation from their sister taxon in Africa, followed by the separation of the European clade and the vicariance of the Asian and North America clades[272,322]. Around 27 kya, the North American barn swallows back-colonized the Russian Baikal region, with the relative subspecies sharing similar coloration than that of the others[272]. Moreover, migratory *H. r. rustica* and the sedentary *H. r. transitiva* show no differentiation, probably due to the intermingling between the two subspecies[322]. Follow up studies on complete mitochondrial genomes, such as Carter and collaborators[324], will be necessary to refine the species phylogeny (see **Chapter 3**).

2.2 The European nightjar

The European nightjar (*Caprimulgus europaeus*) is an insectivorous, crepuscular, ground-nesting bird distributed throughout the Western Palearctic[325], with six recognized subspecies[326]. Its body is 26-28 cm long, with a tail around 10 cm, a wingspan between 57 and 64 cm and a cryptic plumage[327] (**Figure 7a**). European nightjars are also known as "goatsuckers" ("succiacapre" in Italian, "*caprimulgus*" in Latin) according to an ancient folk tale which believed that they feed off goat's milk, which would then stop milk production and eventually go blind. Another belief is that they feed of livestocks blood. Indeed, they feed around livestocks wounds, but just to eat insects attracted to them. Nightjars belong to the Caprimulgidae family, closely related to other bird families (potoos, frogmouths, owlet-nightjars and oilbirds), which together constitute a group of birds (Caprimulgiformes) with peculiar characteristics[328]. For example, the Common Poorwill (*Phalaenoptilus nuttallii*) is the only bird able to hibernate[329–331], while the Oilbird (*Steatornis caripensis*) can echo-localize[332]. These remarkable elusive birds offer many challenges to ornithological studies, including the discovery of new

species[328]. Nightjars have small weak bills, and studies on their cranial conformation revealed that the lower jaw has a specialized spreading mechanism which allows the mouth to open enormously to catch large insects (**Figure 7b**), which, together with their sensitive palate and rictal bristles, may be specializations to nocturnal aerial feeding (see Cleere 2010[328]). Their eyes are large and placed laterally to increase their view during hunting. Moreover, they have the tapetum, a reflective surface in the back of the eye, situated behind a layer of retinal photoreceptors, which increases light detection[328].

Figure 7. Two European nightjars sampled in Ventotene, Latina (Italy) in 2019 during my first PhD year.

The European nightjar breeds in dry, open country with sparse trees and bushes. Wintering areas were found in sub-saharan Africa. The migratory behavior of the European nightjar was only recently being described with the use of geolocators, since ringing efforts provided poor information[326,333,334]. European nightjars experienced a decline, with populations being smaller than those of the late 19th century[335]. Although the European nightjar was at the center of many conservation efforts across Europe (e.g. UK[336], Switzerland[337,338] and Belgium[339]), the species is listed as "Least concern" by the IUCN[340] given its large population size and huge breeding range. Many factors are known to have contributed to its population decline: loss of foraging habitats[338,339], decrease in preys abundance[338], light pollution[341] and, in general, human disturbance[342,343]. However, the lack of knowledge on its biological aspects, like foraging preferences, hindered the identification of accurate conservation plans[344,345]. Deepening the knowledge on the genomics of this cryptic species can surely provide insights into its peculiarities, making conservation efforts more solid. Few genetic studies were performed using mtDNA and nuclear genes[346–349] to decipher the systematics of this species across nightjars, which was before only based on morphological data. To achieve a well resolved phylogeny, multiple genes and multiple species are required[347]. The availability of a reference genome and gene annotation produced as part of this thesis work, will foster studies on the European nightjar biology,

but will also boost the sequencing of other Caprimulgidae species, which will allow the reconstruction of the correct phylogeny of this peculiar family (see **Chapter 4**).

2.3 The lesser kestrel

The lesser kestrel (*Falco naumanni*, **Figure 8**) is a small migratory Afro-Palearctic colonial raptor (order Falconiformes) distributed across southern regions of Eurasia[350]. Its body is 29-32 cm long, with a tail around 11 cm and a wingspan up to 72 cm[351]. Females are bigger and drabber than males, which show a more colorful livery as most raptors[352]. In Europe, lesser kestrels breed in urban areas surrounded by farmlands, and feed mostly on invertebrates (mostly *Coleoptera*)[353]. Wintering areas include sub-saharan Africa, with a small number of individuals wintering in southern Spain[353,354]. They return to the breeding grounds in March and April[355], and pairs form in late May. Lesser kestrels nest in cavities and exploit already existing structures, such as rock holes, ruins and roof tiles[356]. Clutch size varies between 3-5 eggs which are incubated for around 30 days and nestlings fledge 15-20 days later. In the 1950s, the species experienced a steep population decline[357]. Changes in agricultural practices have impacted farmland birds such as the lesser kestrels, which are experiencing a steeper decline than other birds[358,359]. While the species is now considered of a "least concern"[360], it remains of European conservation concern, being listed in Annex 1 of European "Birds Directive" 2009/147/CE ("SPEC 3" for BirdLife International 2017[361]). Several studies were carried on this species focussing on its ecology[355,362–368], conservation[369–375] and genetics[371,376–380]. Genetic studies on the lesser kestrel, which are mostly based on resequencing data, will also be fostered by the generation of a reference genome (see **Chapter 5**). For example, the assessment of the species population structure is currently lacking and the generation of the reference can be a first step towards understanding it.

Figure 8. A male lesser kestrel sampled in Matera, (Italy) in 2016 during my bachelor's degree.

OUTLINE OF THE STUDY

The aim of this thesis work was to exploit the potentials of genomic data in avian species studies at different levels: *de novo* genome assembly, reconstruction of subspecies relationships, comparative genomics, pangenomics, conservation genomics and related biological questions. Special attention was given to the barn swallow (*Hirundo rustica*), its synanthropic behavior and its subspecies phylogeography. The entire work was carried out in the context of several international consortia focused on genome assembly, specifically the Vertebrate Genomes Project (VGP), the Darwin Tree of Life (DToL) and the European Reference Genome Atlas (ERGA). This section is organized according to the manuscripts that resulted from each of these studies.

Chapter 1 reports the flagship paper of the Vertebrate Genomes Project. It focuses on the assembly of near-error free chromosome-level genomes using the new pipeline developed by the VGP (standard pipeline v1.0 and v1.6). The pipelines allow the combination of four different sequencing technologies: PacBio CLR long reads, 10x Linked-Reads, Bionano Optical maps and Hi-C data. Briefly, contigs are assembled, phased and polished from PacBio long reads, generating the principal pseudo-haplotype and the alternative haplotype contigs[8]. Primary contigs are subjected to two steps of purging to improve haplotype separation, and to three scaffolding steps using the other sequencing technologies cited above. The resulting scaffolds are merged with the alternate contigs and the mitogenome and undergo two steps of polishing to increase the base calling and fill the gaps. The final primary and alternate assemblies are manually curated and then annotated. A VGP trio assembly pipeline v1.0-v1.6 is also presented, enabling the near-complete separation of maternal and paternal haplotypes. It is similar to the standard pipeline, but it also includes maternal and paternal Illumina data to enhance haplotype phasing. The minimum metric for genome assemblies were set as follows: 1 Mb contig NG50, 10 Mb scaffold NG50, 90% of the sequence assigned to chromosomes; Q40 average base quality, and haplotypes assembled as completely and correctly as possible. When these metrics are reached, genes, and in general annotation, result more complete.

I personally received a step-by-step training on the VGP standard pipeline v1.0-v1.6 together with the developers, technicians and other volunteers. Using the pipelines, I have worked, or contributed, to the assembly of 5 birds genomes (**Figure 9**): red-crested turaco (*Tauraco erythrolophus*), common yellowthroat (*Geothlypis trichas*), american flamingo (*Phoenicopterus ruber*), rifleman (*Acanthisitta chloris*), barn swallow (*Hirundo rustica*, see **Chapter 2**).

[8]The principal pseudo-haplotype is the best reconstruction of one of the two haplotypes of a species. The alternate haplotype assembly is composed of alternate haplotigs, which are the alleles belonging to the divergent haplotype.

I also generated the lesser kestrel (*Falco naumanni*, **Figure 9**, see **Chapter 5**) trio assembly with the VGP trio assembly pipeline v1.6. While these species were not included in the first VGP flagship paper, which only considered the very first 17 VGP assembled species, they will contribute to the future studies of the VGP. I also performed some evaluation analyses for Chapter 1 work. In particular, I took care of the generation of the Hi-C contact heatmaps for the Extended Data Figure 7.

Figure 9. From left to right: red-crested turaco (Tauraco erythrolophus, photo by DickDaniels, http://theworldbirds.org/), common yellowthroat (Geothlypis trichas, photo by Dan Pancamo), American flamingo (Phoenicopterus ruber, photographed during a trip in Mexico), rifleman (Acanthisitta chloris, photo by digitaltrails), barn swallow (Hirundo rustica, photo by Malene Thyssen), lesser kestrel (Falco naumanni, photo by Sumeet Moghe).

Among the assembled species, I particularly focused on the barn swallow (*Hirundo rustica*). In **Chapter 2** I present the new chromosome-level assembly for the species, which I assembled with the VGP standard pipeline v1.6, the first pangenome for the species and a catalog of genetic markers. I included the barn swallow primary assembly in a comparative genomics analysis, together with six other chromosome-level Passeriformes genomes and the latest chicken (*Gallus gallus*) reference genome. I aligned the assemblies with Cactus[212] in a reference-free way and I exploited the alignment for an evolutionary constraints study. I scanned the genome to find genes under positive selection (accelerated genes) or negative selection (conserved genes). I also included the reference genome in the first pangenome reference for the barn swallow, together with other 5 individuals we sequenced with PacBio HiFi technology, using the Cactus Pangenome Pipeline. Moreover, we used the VGP-quality genome as a reference for a population genetics study using our HiFi data and all publicly available sequencing data for barn swallows' populations and subspecies. We scanned the genome

again to search for signature of selection, but this time with a Linkage Disequilibrium (LD) analysis. Finally, we performed a titration and a phasing experiment on the HiFi data alone to understand and exploit the potential of this innovative sequencing technique.

Chapter 3 reports a companion study resulted from a collaboration with Antonio Torroni's group at the University of Pavia (Italy) that explores the mitogenome relationships between barn swallow subspecies and their demography through time, using phylogenetic and Bayesian analyses. The entire mitogenome is considered, while several previous studies relied only on mtDNA fragments. A total of 411 barn swallow mitogenomes were analyzed, including a reference mitogenome generated from an Italian individual (*H. r. rustica*) using Sanger sequencing, other 4 Italian Sanger-based mitogenomes, 405 barn swallow mitogenomes from 4 subspecies sequenced with NGS and the mitogenome published in Formenti et al., 2019[318].

Chapters 4 and **5** present newly generated chromosome-level genome assemblies for the European nightjar (*Caprimulgus europaeus*) and the lesser kestrel (*Falco naumanni*), respectively. The European nightjar genome was generated in the context of the Darwin Tree of Life project[215] (https://www.darwintreeoflife.org/). I sampled the female individual used for the assembly during my first PhD year on Ventotene island, Latina, Italy (**Figure 10**). The assembly was performed at the Wellcome Sanger Institute using the VGP assembly pipeline v1.6. The paper reports the main statistics of the assembly. Regarding the lesser kestrel, I generated the haplotype-resolved genome using the VGP trio assembly pipeline v1.6 starting from a female individual (offspring) and its parents.

Figure 10. A female European nightjar sampled in Ventotene island, Latina, Italy, in 2019 using heparinized capillary tubes.

Finally, **chapter 6** reports the perspective paper from the ERGA consortium, which I am part of together with my research group at the University of Milan. The paper is a compendium of reference genomes general relevance in biological studies, but most importantly in conservation genomics.

PUBLICATIONS

Chapter 1

Rhie et al. (2021) **“Towards complete and error-free genome assemblies of all vertebrate species”**. *Nature*.

Article

Towards complete and error-free genome assemblies of all vertebrate species

https://doi.org/10.1038/s41586-021-03451-0

Received: 22 May 2020

Accepted: 12 March 2021

Published online: 28 April 2021

Open access

Check for updates

A list of authors and their affiliations appears at the end of the paper.

High-quality and complete reference genome assemblies are fundamental for the application of genomics to biology, disease, and biodiversity conservation. However, such assemblies are available for only a few non-microbial species[1–4]. To address this issue, the international Genome 10K (G10K) consortium[5,6] has worked over a five-year period to evaluate and develop cost-effective methods for assembling highly accurate and nearly complete reference genomes. Here we present lessons learned from generating assemblies for 16 species that represent six major vertebrate lineages. We confirm that long-read sequencing technologies are essential for maximizing genome quality, and that unresolved complex repeats and haplotype heterozygosity are major sources of assembly error when not handled correctly. Our assemblies correct substantial errors, add missing sequence in some of the best historical reference genomes, and reveal biological discoveries. These include the identification of many false gene duplications, increases in gene sizes, chromosome rearrangements that are specific to lineages, a repeated independent chromosome breakpoint in bat genomes, and a canonical GC-rich pattern in protein-coding genes and their regulatory regions. Adopting these lessons, we have embarked on the Vertebrate Genomes Project (VGP), an international effort to generate high-quality, complete reference genomes for all of the roughly 70,000 extant vertebrate species and to help to enable a new era of discovery across the life sciences.

Chromosome-level reference genomes underpin the study of functional, comparative, and population genomics within and across species. The first high-quality genome assemblies of human[1] and other model species (for example, *Caenorhabditis elegans*[2], mouse[3], and zebrafish[4]) were put together using 500–1,000-base pair (bp) Sanger sequencing reads of thousands of hierarchically organized clones with 200–300-kilobase (kb) inserts, and chromosome genetic maps. This approach required tremendous manual effort, software engineering, and cost, in decade-long projects. Whole-genome shotgun approaches simplified the logistics (for example, in human[7] and *Drosophila*[8]), and later next-generation sequencing with shorter (30–150-bp) sequencing reads and short insert sizes (for example, 1 kb) ushered in more affordable and scalable genome sequencing[9]. However, the shorter reads resulted in lower-quality assemblies, fragmented into thousands of pieces, where many genes were missing, truncated, or incorrectly assembled, resulting in annotation and other errors[10]. Such errors can require months of manual effort to correct individual genes and years to correct an entire assembly. Genomic heterozygosity posed additional problems, because homologous haplotypes in a diploid or polyploid genome are forced together into a single consensus by standard assemblers, sometimes creating false gene duplications[11–14].

To address these problems, the G10K consortium[5,6] initiated the Vertebrate Genomes Project (VGP; https://vertebrategenomesproject.org) with the ultimate aim of producing at least one high-quality, near error-free and gapless, chromosome-level, haplotype-phased, and annotated reference genome assembly for each of the 71,657 extant named vertebrate species and using these genomes to address fundamental questions in biology, disease, and biodiversity conservation. Towards this end, having learned the lessons of having too many variables that make conclusions more difficult to reach in the G10K from the G10K Assemblathon 2 effort[15], we first evaluated multiple genome sequencing and assembly approaches extensively on one species, the Anna's hummingbird (*Calypte anna*). We then deployed the best-performing method across sixteen species representing six major vertebrate classes, with a wide diversity of genomic characteristics. Drawing on the principles learned, we improved these methods further, discovered parameters and approaches that work better for species with different genomic characteristics, and made biological discoveries that had not been possible with the previous assemblies.

Complete, accurate assemblies require long reads

We chose a female Anna's hummingbird because it has a relatively small genome (about 1 Gb), is heterogametic (has both Z and W sex chromosomes), and has an annotated reference of the same individual built from short reads[16]. We obtained 12 new sequencing data types, including both short and long reads (80 bp to 100 kb), and long-range linking information (40 kb to more than 100 Mb), generated using eight technologies (Supplementary Table 1). We benchmarked all technologies and assembly algorithms (Supplementary Table 2) in isolation and in many combinations (Supplementary Table 3). To our knowledge, this was the first systematic analysis of many sequence technologies, assembly algorithms, and assembly parameters applied on the same individual. We found that primary contiguous sequences (contigs) (pseudo-haplotype; Supplementary Note 1) assembled from Pacific Biosciences continuous long reads (CLR) or Oxford Nanopore long

reads (ONT) were approximately 30- to 300-fold longer than those assembled from Illumina short reads (SR), regardless of data type combination or assembly algorithm used (Fig. 1a, Supplementary Table 3). The highest contig NG50s for short-read-only assemblies were about 0.025 to 0.169 Mb, whereas for long reads they were about 4.6 to 7.66 Mb (Fig. 1a); contig NG50 is an assembly metric based on a weighted median of the lengths of its gapless sequences relative to the estimated genome size. After fixing a function in the PacBio FALCON software[17] that caused artificial breaks in contigs between stretches of highly homozygous and heterozygous haplotype sequences (Supplementary Note 1, Supplementary Table 2), contig NG50 nearly tripled to 12.77 Mb (Fig. 1a). These findings are consistent with theoretical predictions[18] and demonstrate that, given current sequencing technology and assembly algorithms, it is not possible to achieve high contig continuity with short reads alone, as it is typically impossible to bridge through repeats that are longer than the read length.

Iterative assembly pipeline

Scaffolds generated with all three scaffolding technologies (that is, 10X Genomics linked reads (10XG), Bionano optical maps (Opt.), and Arima Genomics, Dovetail Genomics, or Phase Genomics Hi-C) were approximately 50% to 150% longer than those generated using one or two technologies, regardless of whether we started with short- or long-read-based contigs (Fig. 1b, Extended Data Fig. 1a, Supplementary Table 3). These findings include improvements we made to each approach (Supplementary Note 1, Supplementary Tables 4, 5, Supplementary Fig. 1). Despite similar scaffold continuity, the short-read-only assemblies had from about 18,000 to about 70,000 gaps, whereas the long-read assemblies had substantially fewer (about 400 to about 4,000) gaps (Fig. 1c). Many gaps in the short-read assemblies were in repeat or GC-rich regions. Considering the curated version of this assembly to be more accurate, we also identified roughly 5,000 to 8,000 mis-joins in short-read-based assemblies, whereas long-read-based assemblies had only from 20 to around 700 mis-joins (Fig. 1d). These mis-joins included chimeric joins and inversions. After we curated this assembly for contamination, assembly errors, and Hi-C-based chromosome assignments (Fig. 1e, f), the final hummingbird assembly had 33 scaffolds that closely matched the chromosome karyotype in number (33 of 36 autosomes plus sex chromosomes) and estimated sizes (approximately 2 to 200 Mb; Fig. 1g, h), with only 1 to 30 gaps per autosome (bCalAnn1 in Supplementary Table 6). Of the five autosomes with only one gap each, three (chromosomes 14, 15, and 19) had complete spanning support by at least two technologies (reliable blocks, Extended Data Fig. 1c; bCalAnn1 in Supplementary Table 6), indicating that the chromosome contigs were nearly complete. However, they were missing long arrays of vertebrate telomere repeats within 1 kb of their ends (Extended Data Fig. 1c; bCalAnn1 in Supplementary Tables 6, 7).

Assembly pipeline across vertebrate diversity

Using the formula that gave the highest-quality hummingbird genome, we built an iterative VGP assembly pipeline (v1.0) with haplotype-separated CLR contigs, followed by scaffolding with linked reads, optical maps, and Hi-C, and then gap filling, base call polishing, and finally manual curation (Extended Data Figs. 2a, 3a). We systematically tested our pipeline on 15 additional species spanning all major vertebrate classes: mammals, birds, non-avian reptiles, amphibians, teleost fishes, and a cartilaginous fish (Supplementary Tables 8, 9, Supplementary Note 2). For the zebra finch, we used DNA from the same male as was used to generate the previous reference genome[19], and included a female trio for benchmarking haplotype completeness, where sequenced reads from the parents were used to bin parental haplotype reads from the offspring before assembly[20] (Extended Data Figs. 2a, 3b). We set initial minimum assembly metric goals of: 1 Mb contig NG50; 10 Mb scaffold NG50; assigning 90% of the sequence to chromosomes, structurally validated by at least two independent lines of evidence; Q40 average base quality; and haplotypes assembled as completely and correctly as possible. When these metrics were achieved, most genes were assembled with gapless exon and intron structures[11], and fewer than 3% had frame-shift base errors identified in annotation. Q40 is the mathematical inflection point at which genes go from usually containing an error to usually not[21]. Of the curated assemblies (Supplementary Table 10, Supplementary Note 2), 16 of 17 achieved the desired continuity metrics (Extended Data Table 1). Scaffold NG50 was significantly correlated with genome size (Fig. 2a), suggesting that larger genomes tend to have larger chromosomes. On average, 98.3% of the assembled bases had reliable block NG50s ranging from 2.3 to 40.2 Mb; collapsed repeat bases[22] with abnormally high CLR read coverage (more than 3 s.d.) ranged from 0.7 to 31.4 Mb per Gb; and the completeness of the genome assemblies ranged from 87.2 to 98.1%, with less than 4.9% falsely duplicated regions, consistent with the false duplication rate we found for the conserved BUSCO vertebrate gene set (Extended Data Table 1, Supplementary Tables 11, 12).

Repeats markedly affect continuity

For assemblies generated using our automated pipeline (Extended Data Fig. 3a) before manual curation, all but 2 (the thorny skate and channel bull blenny) of the 17 assemblies exceeded the desired continuity metrics (Supplementary Table 13). In searching for an explanation of these results, we found that contig NG50 decreased exponentially with increasing repeat content, with the thorny skate having the highest repeat content (Fig. 2b, Supplementary Table 13). Consequently, after scaffolding and gap filling, we observed a significant positive correlation between repeat content and number of gaps (Fig. 2c). The kākāpō parrot, which had 15% repeat content, had about 325 gaps per Gb, including 2 of 26 chromosomes with no gaps (chromosomes 16 and 18) and no evidence of collapses or low support, suggesting that the chromosomal contigs were complete (bStrHab1 in Supplementary Table 6). By contrast, the thorny skate, with 54% repeat content, had about 1,400 gaps per Gb (Extended Data Table 1); none of its 49 chromosomal-level scaffolds contained fewer than eight gaps, and all had some regions that contained collapses or low support (sAmbRad1 in Supplementary Table 6). Even after curation and other modifications to increase assembly quality (Supplementary Note 2), the number of collapses, their total size, missing bases, and the number of genes in the collapses all correlated with repeat content (Extended Data Fig. 4a–d). The average collapsed length, however, correlated with average CLR read lengths (10–35 kb; Extended Data Fig. 4e). There were no correlations between the number of collapsed bases and heterozygosity or genome size (Extended Data Fig. 4f, g). Depending on species, 77.4 to 99.2% of the collapsed regions consisted of unresolved segmental duplications (Extended Data Fig. 4h). The remainder were high-copy repeats, mostly of previously unknown types (Extended Data Fig. 4i), and of known types such as satellite arrays, simple repeats, long terminal repeats (LTRs), and short and long interspersed nuclear elements (SINEs and LINEs), depending on species (Extended Data Fig. 4j). We found that repeat masking before generating contigs prevented some repeats from making it into the final assembly (Supplementary Note 3). All of the above findings quantitatively demonstrate the effect that repeat content has on the ability to produce highly continuous and complete assemblies.

Detection and removal of false duplications

During curation, we discovered that one of the most common assembly errors was the introduction of false duplications, which can be misinterpreted as exon, whole-gene, or large segmental duplications.

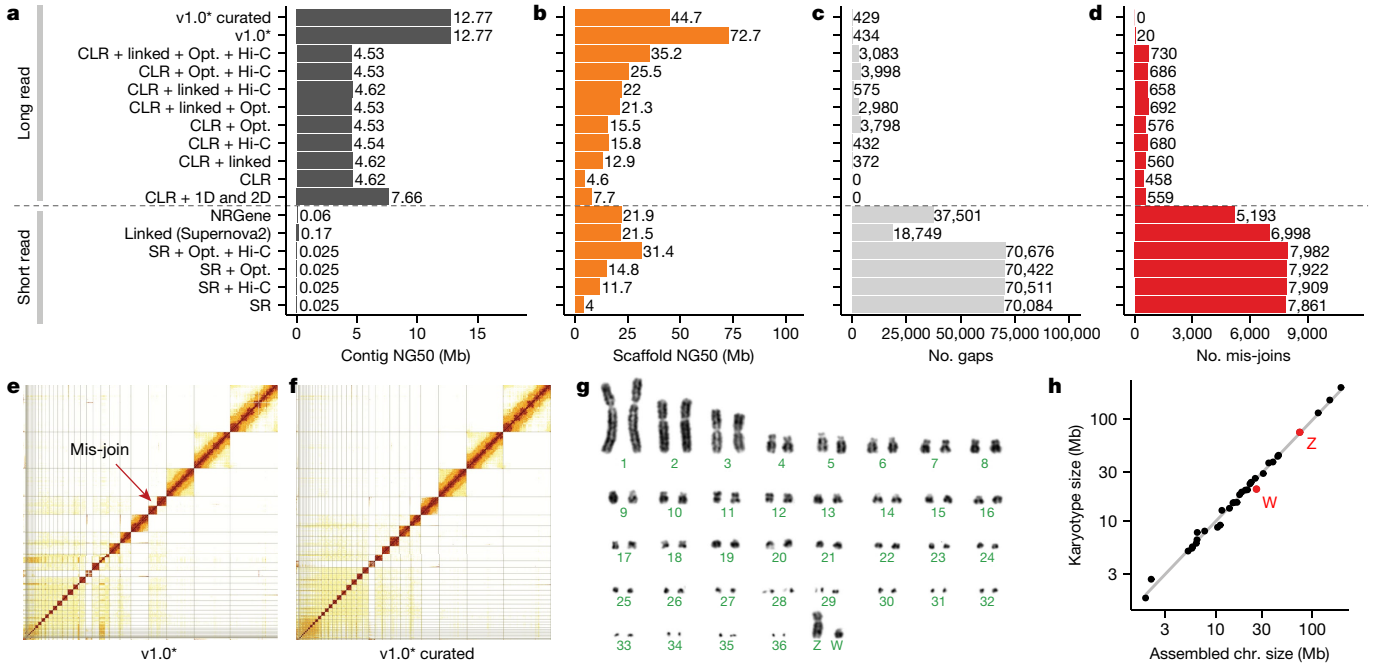


Fig. 1 | Comparative analyses of Anna's hummingbird genome assemblies with various data types. a, Contig NG50 values of the primary pseudo-haplotype. **b**, Scaffold NG50 values. **c**, Number of joins (gaps). **d**, Number of mis-join errors compared with the curated assembly. The curated assembly has no remaining conflicts with the raw data and thus no known mis-joins. *Same as CLR + linked + Opt. + Hi-C, but with contigs generated with an updated FALCON[17] version and earlier Hi-C Salsa version (v2.0 versus v2.2; Supplementary Table 2) for less aggressive contig joining. **e**, **f**, Hi-C interaction heat maps before and after manual curation, which identified 34 chromosomes. Grid lines indicate scaffold boundaries. Red arrow, example mis-join that was corrected during curation. **g**, Karyotype of the identified chromosomes ($n = 36 + ZW$), consistent with previous findings[70]. **h**, Correlation between estimated chromosome sizes (in Mb) based on karyotype images in **g** and assembled scaffolds in Supplementary Table 4 (bCalAna1) on a log–log scale. v1.0, VGP assembly v1.0 pipeline; linked, 10X Genomics linked reads; Hi-C, Hi-C proximity ligation; 1D, 2D, Oxford Nanopore long reads; NRGene, NRGene paired-end Illumina reads; SR, paired-end Illumina short reads.

We observed two types of false duplication: 1) heterotype duplications, which occurred in regions of increased sequence divergence between paternal and maternal haplotypes, where separate haplotype contigs were incorrectly placed in the primary assembly (Extended Data Fig. 5a); and 2) homotype duplications, which occurred near contig boundaries or under-collapsed sequences caused by sequencing errors (Extended Data Fig. 5b). False heterotype duplications appeared to occur with higher heterozygosity. For example, during curation of the female zebra finch genome, we found an approximately 1-Mb falsely duplicated heterozygous sequence (Extended Data Fig. 6a). This zebra finch individual had the highest heterozygosity (1.6%) relative to all other genomes (0.1–1.1%). Homotype duplications often occurred at contig boundaries, and were approximately the same length as the sequence reads (Extended Data Fig. 6b, c). We identified and removed false duplications during curation using read coverage, self-, transcript-, optical map- and Hi-C-alignments, and *k*-mer profiles (Extended Data Fig. 6, Supplementary Fig. 2).

Before we purged false duplications, the primary assembly genome size correlated positively with estimated percentage heterozygosity; more heterozygous genomes tended to have assembly sizes bigger than the estimated haploid genome size (Fig. 2d). Similarly, the extra duplication rate in the primary assembly, measured using *k*-mers[23] or conserved vertebrate BUSCO genes[24], varied from 0.3% to 30% and trended towards correlation with heterozygosity (Fig. 2e, f, Supplementary Table 13). Apparent false gene duplication rates correlated more strongly with the overall repeat rate in the assemblies (Fig. 2g, h). To remove these false duplications automatically, we initially used Purge_Haplotigs[13], which removed retained falsely duplicated contigs that were not scaffolded (Extended Data Fig. 5; VGP v1.0–1.5). Later, we developed Purge_Dups[14] to remove both falsely retained contigs and end-to-end duplicated contigs within scaffolds (Extended Data Fig. 5; VGP v1.6), which reduced the amount of manual curation. After we applied these tools, the primary assembly sizes and the *k*-mer and BUSCO gene duplication rates were all reduced, and their correlations with heterozygosity and repeat content were also reduced or eliminated (Fig. 2d–h). These findings indicate that it is essential to properly phase haplotypes and to obtain high consensus sequence accuracy in order to prevent false duplications and associated biologically false conclusions.

Curation is needed for a high-quality reference

Each automated scaffolding method introduced tens to thousands of unique joins and breaks in contigs or scaffolds (Supplementary Table 14). Depending on species, the first scaffolding step with linked reads introduced about 50–900 joins between CLR-generated contigs. Optical maps introduced a further roughly 30–3,500 joins, followed by Hi-C with about 30–700 more joins, and each identified up to several dozen joins that were inconsistent with the previous scaffolding step. Manual curation resulted in an additional 7,262 total interventions for 19 genome assemblies or 236 interventions per Gb of sequence (Supplementary Table 15). When a genome assembly was available for the same or a closely related species, it was used to confirm putative chromosomal breakpoints or rearrangements (Supplementary Table 15). These interventions indicate that even with current state-of-the-art assembly algorithms, curation is essential for completing high-quality reference assemblies and for providing iterative feedback to improve assembly algorithms. A further description of our curation approach and analyses of VGP genomes are presented elsewhere[25].

Hi-C scaffolding and cytological mapping

Most large assembled scaffolds of each species spanned entire chromosomes, as shown by the relatively clean Hi-C heat map plots across each

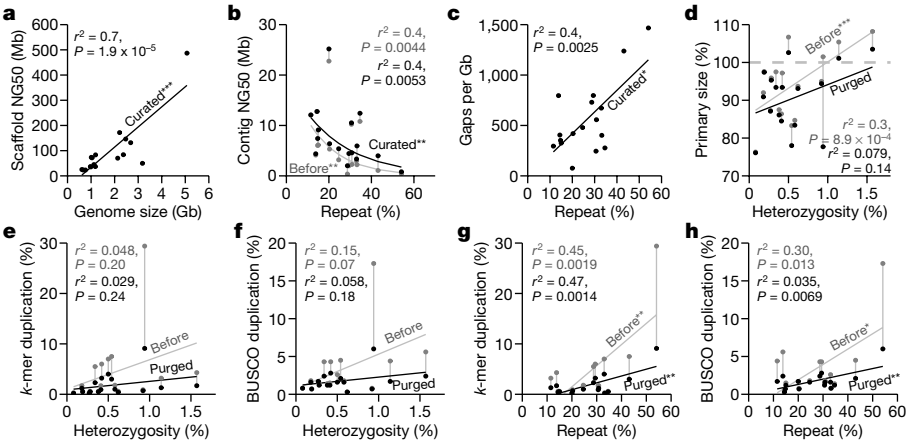


Fig. 2 | Impact of repeats and heterozygosity on assembly quality. **a**, Correlation between scaffold NG50 and genome size of the curated assemblies. **b**, Nonlinear correlation between contig NG50 and repeat content, before and after curation. **c**, Correlation between number of gaps per Gb assembled and repeat content. **d**, Correlation between primary assembly size relative to estimated genome size (*y* axis) and genome heterozygosity (*x* axis), before and after purging of false duplications. Assembly sizes above 100% indicate the presence of false duplications and those below 100% indicate collapsed repeats. **e**, **f**, Correlations between genome duplication rate using *k*-mers[23] (**e**) and conserved BUSCO vertebrate gene set (**f**), and genome heterozygosity before and after purging of false duplications. **g**, **h**, As in **e**, **f**, but with whole-genome repeat content before and after purging of false duplications. Genome size, heterozygosity, and repeat content were estimated from 31-mer counts using GenomeScope[71], except for the channel bull blenny, as the estimates were unreliable (see Methods). Repeat content was measured by modelling the *k*-mer multiplicity from sequencing reads. Sequence duplication rates were estimated with Merqury[23] using 21-mers. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$, of the correlation coefficient: *P* values and adjusted r^2 from *F*-statistics. $n = 17$ assemblies of 16 species.

scaffold after curation (Extended Data Fig. 7), near perfect correlation between chromosomal scaffold length and karyotypically determined chromosome length (Fig. 1h), and the presence of telomeric repeat motifs on some scaffold ends (Supplementary Table 7). In our VGP zebra finch assembly, all inferred chromosomes were consistent with previously identified linkage groups in the Sanger-based reference, except for chromosomes 1 and 1B (Extended Data Fig. 8a). Their join in the VGP assembly was supported by both single CLR reads and optical maps through the junction. We also corrected nine inversion errors and filled in large gaps at some chromosome ends. In the platypus, we identified 18 structural differences in 13 scaffolds between the VGP assembly and the previous Sanger-based reference anchored to chromosomes using fluorescence in situ hybridization (FISH) physical mapping (Extended Data Fig. 8b, Supplementary Table 16). Of these 18, all were supported with Hi-C, and seven were also supported by both CLR and optical maps in the VGP assembly. Our platypus assembly also filled in many large (approximately 1–30 Mb) gaps and corrected many inversion errors (Extended Data Fig. 8b). Furthermore, we identified seven additional chromosomes (chromosomes 30–36) in the zebra finch, and eight (chromosomes 8, 9, 14, 15, 17, 19, 21, and X4; Extended Data Fig. 8a, b) in the platypus[26,27]. Relative to the VGP assembly, the earlier short-read Anna's hummingbird assembly was highly fragmented (Extended Data Fig. 8c), despite being scaffolded with seven different Illumina libraries spanning a wide range of insert sizes (0.2–20 kb). The previous climbing perch assembled chromosomes were even more fragmented and also had large gaps of missing sequence (Extended Data Fig. 8d). On average, 97% ± 3% (s.d.) of the assembled bases were assigned to chromosomes (Extended Data Table 1), compared with 76% and 32% in the prior zebra finch and platypus references, respectively. We believe the comparable or higher accuracy of Hi-C relative to genetic linkage or FISH physical mapping is due to the higher sampling rate of Hi-C pairs across the genome. Nonetheless, visual karyotyping is useful for complementary validation of chromosome count and structure[28].

Trios help to resolve haplotypes

We were able to assemble the trio-based female zebra finch contigs into separate maternal and paternal chromosome-level scaffolds (Extended Data Fig. 9a) using our VGP trio pipeline (Extended Data Fig. 3b). Compared to the non-trio assembly of the same individual, the trio version had seven- to eightfold fewer false duplications (*k*-mer and BUSCO dups in Supplementary Tables 11, 12), well-preserved haplotype-specific variants (*k*-mer precision/recall 99.99/97.08%), and higher base call accuracy, exceeding Q43 for both haplotypes (Extended Data Table 1). The trio-based assembly was the only assembly with nearly perfect (99.99%) separation of maternal and paternal haplotypes, determined using *k*-mers specific to each[23]. We identified haplotype-specific structural variants, including inversions of 4.5 to 12.5 Mb on chromosomes 5, 11, and 13 that were not readily identifiable in the non-trio version (Extended Data Fig. 10a–e). Moving forward, the VGP is prioritising the collection of mother–father–offspring trios where possible, or single parent–offspring duos, to assist with diploid assembly and phasing, as well as the development of improved methods for the assembly of diploid genomes in the absence of parental genomic data, as described in another study[29].

Effects of polishing on accuracy

Despite their increased continuity and structural accuracy, CLR-based assemblies required at least two rounds of short-read consensus polishing to reach 99.99% base-level accuracy (one error per 10 kb, Phred[30] Q40; Supplementary Table 5). Before polishing, the per-base accuracy was Q30–35 (calculated using *k*-mers). The most common errors were short indels from inaccurate consensus calling during CLR contig formation, which resulted in amino acid frameshift errors. Using our combined approach of long-read and short-read polishing applied on both primary and alternate haplotype sequences together, we polished from 82% to 99.7% of the primary and about 91.3% of the alternate assembly (Supplementary Table 17). Of the remaining unpolished sequence, one haplotype was sometimes reconstructed at substantially lower quality, because most reads aligned to the higher quality haplotype (Extended Data Fig. 11a). False duplications had similar effects, where the duplicated sequence acted as an attractor during the read mapping. Haplotypes in the more homozygous regions tended to be collapsed by FALCON-Unzip[17]. All such cases recruited reads from both haplotypes and thereby caused switch errors, which we confirmed in the trio-based assembly and fixed when excluding read pairs from the other haplotype

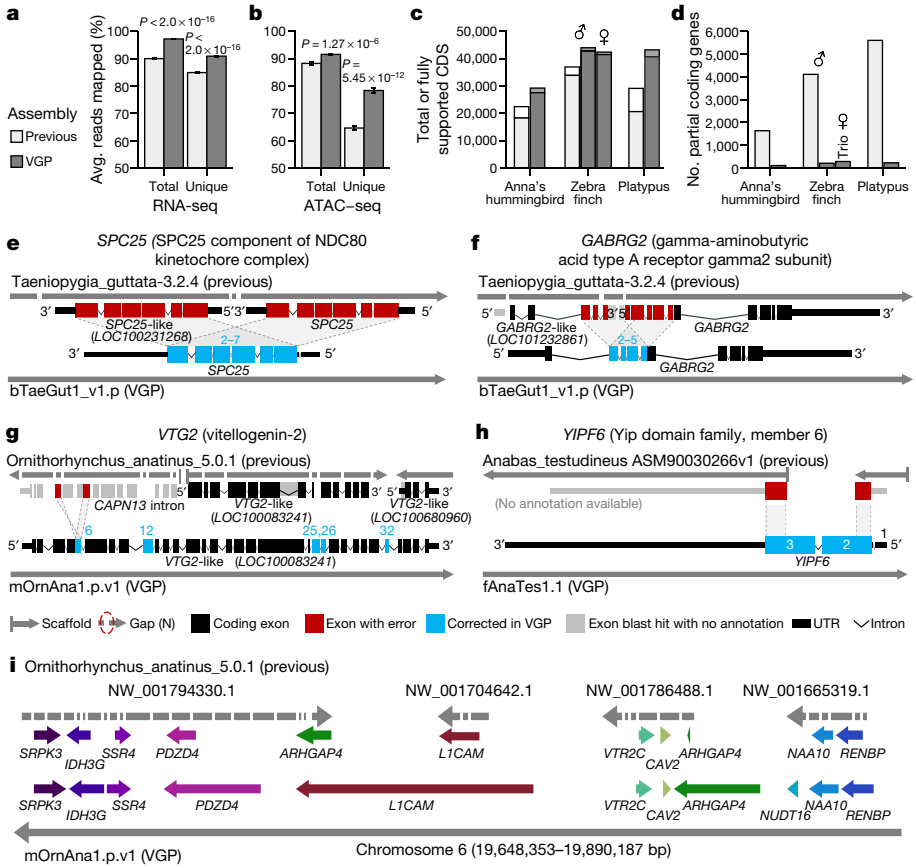


Fig. 3 | Improvements to alignments and annotations in VGP assemblies relative to prior references. a, **b**, Average percentage of RNA-seq transcriptome samples (**a**; $n = 44$, mean ± s.e.m.) and ATAC–seq genome reads (**b**; $n = 12$) that align to the previous and VGP zebra finch assemblies. Unique reads mapped to only one location in the assembly. Total is the sum of unique and multi-mapped reads. *P* values are from paired *t*-test. **c**, **d**, Total number of coding sequence (CDS) transcripts (full bar) and portion fully supported (inner bar) (**c**) and the number of RefSeq coding genes annotated as partial (**d**) in the previous and VGP assemblies using the same input data. **e**–**h**, Examples of assembly and associated annotation errors in previous reference assemblies corrected in the new VGP assemblies. See main text for descriptions. **i**, Gene synteny around the *VTR2C* receptor in the platypus shows completely missing genes (*NUDT16*), truncated and duplicated *ARHGAP4*, and many gaps in the earlier Sanger-based assembly compared with the filled in and expanded gene lengths in the new VGP assembly. Assembly accessions are in Supplementary Table 19.

during polishing (Extended Data Fig. 11b). These findings indicate that both sequence read accuracy and careful haplotype separation are important for producing accurate assemblies.

Sex chromosomes and mitochondrial genomes

Sex chromosomes have been notoriously difficult to assemble, owing to their greater divergence relative to autosomes and high repeat content[31]. We successfully assembled both sex chromosomes (Z, W) for all three avian species, the first W chromosome (to our knowledge) for vocal learning birds (Extended Data Figs. 7, 9b), the X and/or Y chromosome in placental mammals (Canada lynx and two bat species), the X chromosome in the thorny skate, and for the first time, to our knowledge, all ten sex chromosomes (5X and 5Y) in the platypus[26] (Extended Data Fig. 9c). The completeness and continuity of the zebra finch Z and W chromosomes were further improved by the trio-based assembly (Extended Data Fig. 9b). However, the sex chromosome assemblies were still more fragmented than the autosomes, probably owing to their lower sequencing depth and high repeat content.

Mitochondrial (MT) genomes, which are expected to be 11–28 kb in size[32], were initially found in only six assemblies (Supplementary Table 18). The MT-derived raw reads were present, but they failed to assemble, in part because of minimum read-length cutoffs for the starting contig assembly. Furthermore, if the MT genome was not present during nuclear genome polishing, the raw MT reads were attracted to nuclear MT sequences (NuMTs), incorrectly converting them to the full organelle MT sequence (Extended Data Fig. 11c). To address these issues, we developed a reference-guided MT pipeline and included the MT genome during polishing[33] (Extended Data Fig. 3c; VGP v1.6). With these improvements, we reliably assembled 16 of 17 MT genomes (Supplementary Table 18) and discovered 2 kb of an 83-bp repeat expansion within the control region in the kākāpō (Extended Data Fig. 9d), and *Nad1* and *trnL2* gene duplications in the climbing perch (Extended Data Fig. 9e). These duplications were verified using single-molecule CLR reads that spanned the duplication junctions or even the entire MT genome. Their absence in previous MT references[34,35] is likely to result from the inability of Sanger or short reads to correctly resolve large duplications. More details on the MT-VGP pipeline and new biological discoveries are reported elsewhere[33].

Improvements to read alignment and annotation

Compared to previous Sanger (zebra finch and platypus) and Illumina (Anna's hummingbird and climbing perch) assemblies, we added about 42–176 Mb of missing sequence and placed 68.5 Mb (zebra finch) to 1.8 Gb (platypus) of previously unplaced sequence within chromosomes. We corrected about 7,800–64,000 mis-joins, and closed 55,177–193,137 gaps per genome (Supplementary Table 19). Consistent with these improvements, both transcriptome RNA sequencing (RNA-seq) data (Fig. 3a) and genome assay for transposase-accessible chromatin using sequencing (ATAC–seq) data (Fig. 3b) aligned with about 5 to 10% greater mapability to our new VGP assemblies compared with the previous assemblies. The NCBI RefSeq and EBI Ensembl annotations revealed: 5,434 to 14,073 more protein-coding transcripts per species, with 94.1 to 97.8% fully supported (Fig. 3c, Supplementary Table 20); only about 100 to 300 partially assembled coding genes, compared with about 1,600 to 5,600 (Fig. 3d); more orthologous coding genes shared with human; and fewer transcripts that required corrections to compensate for premature stop codons or frame-shift indel errors (Extended Data Table 2). The total number of genes annotated went down in the VGP assemblies (Extended Data Table 2), partly because there were fewer false duplications (Supplementary Table 19). Supporting these results, the VGP assemblies had 0 to 13% higher *k*-mer completeness (95% mean ± 3.5% s.d. versus 88 ± 4.3%; Extended Data Table 2, Supplementary Table 19; $P = 0.0047$, $n = 4$ prior and 17 VGP assemblies, unpaired *t*-test).

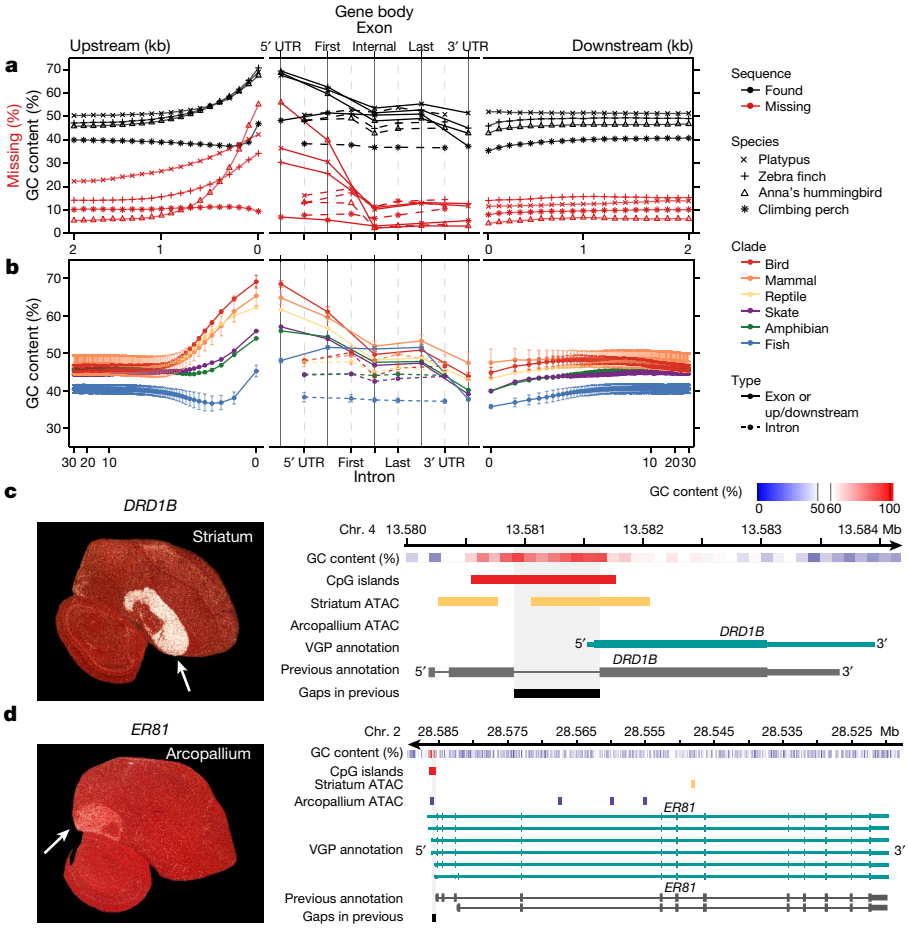


Fig. 4 | VGP assemblies reveal GC content patterns in protein-coding genes. a, Average GC content (n = 14,000–18,000 annotated coding genes; Extended Data Table 2) in VGP assemblies (black) and the percentage of genes with missing sequence in the earlier references (red) based on a Cactus alignment, in 100-bp blocks, 2 kb on either side of all protein-coding genes (left and right), and for UTRs, exons, and introns (middle). **b**, Average GC content (mean ± s.d. for lineages with more than one species) of the six major vertebrate lineages sequenced, for 30 kb upstream and downstream (in 100-bp blocks, log scale; left and right) and of the UTR, exons, and introns (middle). **c**, **d**, Left, specialized expression (arrows) shown by in situ hybridization of *DRD1B* in the zebra finch striatum (**c**) and *ER81* in the arcopallium (**d**), from Jarvis et al.[47]; the cerebellum was removed from the ER81 image. Right, ATAC–seq profiles in the GC-rich promoter regions of these genes, showing each gene's GC content (red is high), the ATAC–seq peaks in striatum (purple) or arcopallium (yellow) neurons, and portions of missing sequence (black) in the previous reference assembly (grey).

An example of a whole-gene heterotype false duplication in the RefSeq annotation of the previous zebra finch reference[19] is the BUSCO gene *SPC25*[36], for which each haplotype was correctly placed in the VGP primary and alternate assemblies (Fig. 3e). The *GABRG2* receptor, which shows specialized expression in vocal learning circuits[37], had a partial tandem duplication of four of its ten exons, resulting in annotated partial false tandem gene duplications (*GABRG2* and *GABRG2*-like; Fig. 3f). The vitellogenin-2 (*VTG2*) gene, a component of egg yolk in all egg-laying species[38], was distributed across 14 contigs in 3 different scaffolds in the previous platypus assembly (Fig. 3g). Two of these scaffolds received two corresponding *VTG2*-like gene annotations, and the third was included as false duplicated intron in *CAPN-13* (red), together causing false amino acid sequences in five exons (blue). The BUSCO *YIPF6* gene, which is associated with inflammatory bowel disease[39], was split between two different scaffolds and is thus presumed to be a gene loss in the earlier climbing perch assembly[40] (Fig. 3h). Each of these genes is now present on long VGP contigs, within validated blocks, with no gaps and no false gene gains or losses (Supplementary Table 21).

Going beyond individual genes, a ten-gene synteny window surrounding the vasotocin receptor 2C gene (*VTR2C*; also known as *AVPR2*), which is involved in blood pressure homeostasis and brain function[41,42], was split into 34 contigs on four scaffolds, one of which contained a false haplotype duplication of *ARHGAP4* in the previous platypus assembly[43] (Fig. 3i). In our VGP assembly, all eleven genes were in one 37-Mb-long contig within the approximately 50 Mb chromosome 6 scaffold. Furthermore, eight of the eleven genes were remarkably increased in size owing to the addition of previously unknown missing sequences. This chromosomal region was more GC-rich (54%) than the entire chromosome 6 (46%). Thousands of such false gains and losses in previous reference assemblies have been corrected in our VGP assemblies (more details in refs. [27,44]), demonstrating that assembly quality has a critical effect on subsequent annotations and functional genomics.

GC-rich regulatory regions of coding genes

We tested whether the higher-quality VGP assemblies enabled new biological discoveries. Notably, beginning about 1.5 kb upstream of protein-coding genes, in 100-bp blocks, there was a steady increase from about 6–20% to about 30–55% of genes having missing sequence in previous references (Fig. 4a); similarly high proportions of genes were missing their subsequent 5′ untranslated regions (UTRs) and first exons. This fluctuation in missing sequence was directly proportional to GC content (Fig. 4a). We therefore studied the GC content pattern across all protein-coding genes in all 16 new VGP assemblies and found a genome-wide signature: a rapid rise in GC content in the roughly 1.5 kb before the transcription start site, in the 5′ UTR, and in the first exon, followed by a steady decrease in subsequent exons and returning to near intergenic background levels in the 3′ UTR and about 1.5 kb after the transcription termination site (Fig. 4b). The introns had lower GC content, closer to the intergenic background. The intergenic GC content was stable within 30 kb on either side of each gene (Fig. 4b). Mammals, birds, and reptiles had the highest increase (around 20%) in GC content near the start site, followed by the amphibian and skate with medium levels (around 10%). Teleost fishes showed an initial decrease, followed by weaker increase (about 5%) from an already lower GC content (Fig. 4b). Given that the skate represents the sister branch to all other vertebrate lineages sequenced, these findings suggest that teleosts lost at least 5% GC content genome-wide, while maintaining most of the GC content pattern in protein-coding genes. Although it is known that promoter regions can be CpG rich, and GC content can vary between

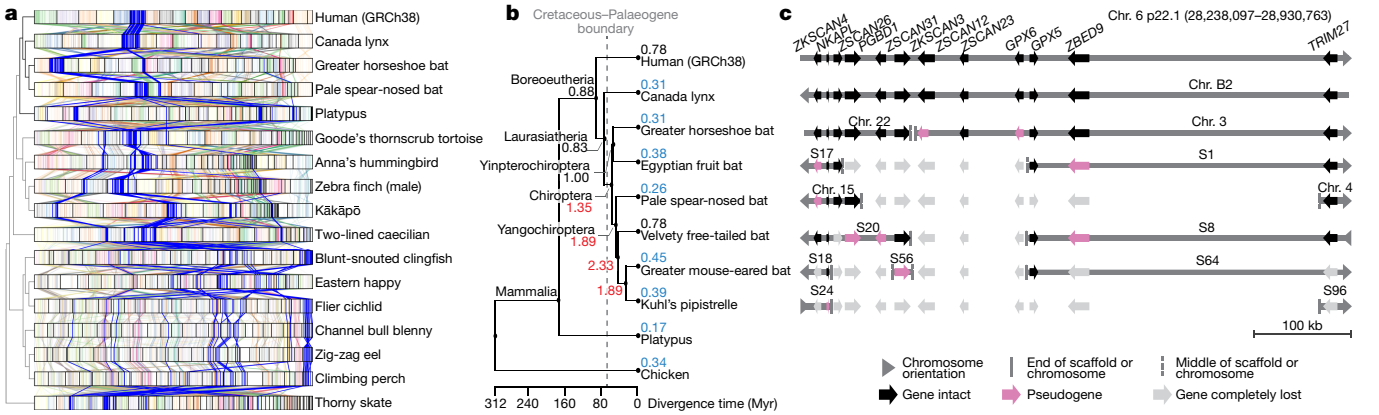


Fig. 5 | Chromosome evolution among bats and other vertebrates. **a**, Chromosome synteny maps across the species sequenced based on BUSCO gene alignments. Chromosome sizes (bar lengths) are normalized to genome size, to make visualization easier. Genes (lines) are coloured according to the human chromosome to which they belong; those on human chromosome 6 are highlighted in blue and other chromosomes are in lighter shades. The cladogram is from the TimeTree database[72]. **b**, Phylogenetic relationship of the mammalian species sequenced and their inferred chromosome EBR rates (breaks per Myr) on different branches. Red, higher rates than average (0.84); blue, lower than average. **c**, Summary of alignment, gene organization, and functional gene status surrounding a bat interchromosomal EBR involving the homologue of human chromosome 6. End of scaffold (S) or chromosome (Chr.) means that the breakpoint is located at a chromosome arm end; middle means that it is located within a scaffold or chromosome. Scale is relevant for human Chr. 6 only. Actual gene sizes in the non-human species may differ and were drawn to match the annotated human gene sizes for simplicity.

exons and introns[45,46], such a systematic pattern, the lineage-specific differences within vertebrates, and the magnitude of these differences had not been previously described, to our knowledge.

We tested whether the newly assembled GC-rich promoter regions contained novel regulatory sequences. Analysing the zebra finch brain, we found that genes with upregulated expression specific to the striatum (for example, *DRD1B*, which encodes a dopamine receptor) had ATAC-seq peaks in the GC-rich promoter and 5′ UTR region in striatal neurons, but not in arcopallium neurons (Fig. 4c); conversely, genes (for example, the *ER81* transcription factor) with upregulated expression in the arcopallium (mammalian cortex layer 5 equivalent[47]) had ATAC–seq peaks in the GC-rich region in arcopallium neurons but not striatal neurons (Fig. 4d). These GC-rich regions were missing in the earlier assembly. In addition, the missing region in *DRD1B* led to a false annotation as a two-exon gene[48], whereas the VGP assembly revealed a single-exon gene (Fig. 4c). These GC-rich promoter regions are candidates for driving cell-type-specific expression. These findings demonstrate the importance of using sequencing chemistry that reads through GC-rich regions, like the CLR method. The earlier hummingbird genome assembly was generated using Illumina TruSeq3 chemistry[16], which was designed to read through GC-rich regions, and yet about 55% of the genes were missing the 100-bp GC-rich region before the start site (Fig. 4a). Another paper contains additional findings on missing regions[27].

Chromosomal evolution

We next investigated whether we could gain new insights into chromosome evolution among vertebrates. Given the more than 430 million years (Myr) of evolutionary divergence among the species sampled here, it was difficult to generate whole genome-to-genome alignments across all species. Thus, we focused our initial analyses on 1,147 highly conserved BUSCO vertebrate genes that are shared among our assemblies of all 16 species and the human reference (GRCh38). Human chromosomes mapped with greater orthology to 3.7 ± 1.3 (s.d.) chromosomes on average in other mammals, compared to 5.6 ± 2.2 in amphibians and 9.6 ± 3.3 in teleost fishes (Fig. 5a, Supplementary Table 22). The skate chromosome arrangement was more conserved with tetrapods, mapping to 2.9 ± 1.4 chromosomes on average, compared to 4.8 ± 2.5 in teleost fishes. These findings indicate that, along with a reduction in GC content, the teleost lineage has experienced more massive chromosome rearrangements since divergence from their most recent common ancestor with tetrapods, consistent with a proposed higher rearrangement rate in teleosts[49].

To determine the precise locations of chromosome rearrangements between species, we focused on a shorter evolutionary distance of around 180 Myr among mammals, and added four additional bat species described in our Bat1K study[50], the human genome reference[51] (GRCh38.p12), and a recently upgraded long-read chicken reference[52] (galGal6a) as an outgroup. Pairwise whole-genome alignments to the human reference defined homologous synteny blocks and evolutionary breakpoint regions (EBRs) among the species. We found that breakpoint rates (EBRs per Myr) tripled among bats soon after the last mass extinction event (about 66 million years ago (Mya)), a time of rapid bat superfamily divergences[53] (about 60 Mya; Fig. 5b). Some rearrangements affected genes. For example, a 1.3-Mb inversion in greater horseshoe bat chromosome 28 (homologous to 29.5 Mb of human chromosome 15; Extended Data Fig. 12a) disrupted *STARD5*, a gene involved in cholesterol homeostasis in liver cells[54]. The rearrangement separated exons 1–5 from exon 6, and disrupted splicing of the transcripts (Extended Data Fig. 12b). Another example was an EBR that involved fission of an ancestral bat chromosome homologue of human chromosome 6 (boreoeutherian mammal chromosome 5[55]) and was later reused among the different bat lineages in rearrangements that involved the ancestral homologues of human chromosomes 1, 2 and 6 (Fig. 5c, Extended Data Fig. 12c). We also noted a fission in this region in the mouse, rat, and dog genomes[55]. On the basis of the conserved gene order in human and Canada lynx, we inferred that the boreoeutherian ancestral mammal locus corresponding to human 6p22.1 contained 12 genes, including four *ZSCAN* and two *ZKSCAN* transcription factors, and two *GPX* enzyme genes, all associated with sequentially increasing independent gene losses in bats (Fig. 5d). For example, the greater horseshoe bat lost only *ZSCAN12* and *GPX6* to pseudogenization, whereas Kuhl's pipistrelle lost all 12 genes. *ZSCAN* and *ZKSCAN* are involved in cell differentiation, migration and invasion, proliferation, apoptosis, and innate immunity[56]. We speculate that loss of *ZSCAN12* in all six bats could contribute to their immune tolerance to pathogens[50].

Other biological findings using these VGP assemblies are published elsewhere, and include: 1) more accurate synteny across species, leading to a better understanding of the evolution of and thus a universal nomenclature for the vasotocin (also known as vasopressin) and

Table 1 | Proposed standards and metrics for defining genome assembly quality

Quality category	Metric	Finished	VGP-2020	VGP-2016	B10k-2014	This study
Notation	***x.y.P.Q.C***	c.c.Pc.Q60.C100	7.c.P6.Q50.C95	6.7.P5.Q40.C90	4.5.Q30	
Continuity	**Contig NG50 (*x*)**	= Chr. NG50	>10 Mb	>1 Mb	>10 kb	1–25 Mb
	Scaffolds NG50 (*y*)	= Chr. NG50	= Chr. NG50	>10 Mb	>100 kb	23–480 Mb
	Gaps per Gb	No gaps	<200	<1,000	<10,000	75–1,500
Structural accuracy	**Reliable blocks**	= Chr. NG50	>10 Mb	>1 Mb	Not required	2.3–40.2 Mb
	False duplications	0%	<1%	<5%	<10%	0.2–5.0%
	Curation	Conflicts resolved	Manual	Manual	Not required	Manual
Base accuracy	**Base pair QV (*Q*)**	>60	>50	>40	>30	39–43
	***k*-mer completeness**	100% complete	>95%	>90%	>80%	87–98%
Haplotype phasing	**Phase block NG50 (*P*)**	= Chr. NG50	>1 Mb	>100 kb	Not required	1.6 Mb[a]
Functional completeness	**Genes**	>98% complete	>95% complete	>90%	>80%	82–98%
	Transcript mappability	>98%	>90%	>80%	>70%	96%
Chromosome status	**Assigned (C)**	>100%	>95%	>90%	Not required	94.4–99.9%
	Sex chromosomes	Right order, no gaps	Localized homo pairs	At least one shared (for example, X or Z)	Fragmented	At least one shared
	Organelles (for example, MT)	One complete allele	One complete allele	Fragmented	Not required	One complete allele

The six broad quality categories in the first column are split into sub-metrics in the second column. The recommendations for draft to finished qualities (columns 3–6) are based on those achieved in past studies[16,19,63], this study, and what we aspire to. In the *x.y.P.Q.C* notation, $x = \log_{10}$[contig NG50]; $y = \log_{10}$[scaffold NG50]; $P = \log_{10}$[haplotype phased NG50 block]; Q = Phred base accuracy QV; and C = percentage of the assembly assigned to chromosomes. c denotes 'complete' telomere-to-telomere continuity. The VGP assemblies (last column) satisfy the 6.7.6.Q40.C90 standard, but some come close to achieving a higher 7.c.7.Q50.C95 standard. These metrics apply to genomes about 1 Gb or bigger.

[a]Phase blocks calculated for the zebra finch non-trio assembly using haplotype specific *k*-mers from parental data[20]; the trio assemblies had NG50 phase blocks of 17.3 Mb (maternal) and 56.6 Mb (paternal).

oxytocin ligand and receptor gene families[57]; 2) greater understanding of the evolution of the carbohydrate 6-*O* sulfotransferase gene family, which encodes enzymes that modify secreted carbohydrates[58]; 3) the first Bat1K study[50], which generated a genome-scale phylogeny that better resolves the relationships between bats and other mammals, and which identified changes in bat genes that are involved in immunity and life span, including genes that are relevant to the COVID-19 pandemic[59]; 4) deleterious mutations that have been purged from the last surviving isolated and inbred population of the critically endangered kākāpō[60]; and 5) more complete resolution of the evolution of the complex sex chromosomes in platypus and echidna[26]. These discoveries were not possible with the previous reference assemblies, and we expect many future discoveries to follow.

Proposed assembly quality metrics

Drawing on the lessons learned from this work, we propose that assembly quality should be summarized using 14 metrics under 6 categories (Table 1; full details in Supplementary Note 4). We summarize the most critical and commonly used metrics using the simple notation *x.y.P.Q.C*, where: $x = \log_{10}$[contig NG50], $y = \log_{10}$[scaffold NG50], $P = \log_{10}$[haplotype phase block NG50], Q = QV base accuracy, and C = percentage of the assembly assigned to chromosomes (Table 1). Our current minimum VGP standard, for example, is 6.7.P5.Q40.C90. This revises our prior notation[50,61,62], which reported log-scaled continuity measured in 'kilobases' rather than 'bases'. The thresholds we chose were based on empirical and quantitative observations between what is achievable currently and what is aspirational, and the question the assemblies are meant to answer. For example, the short-read paired-end library-based assemblies of the B10K Phase 1 genomes in 2014[16] and the 10XG linked-read assembly of the Anna's hummingbird presented here would be categorized as a 4.5.P7.Q50 assembly, with low continuity but high base accuracy (Table 1). Such a genome would be suitable for use in phylogenomics[63] and for population-scale SNP surveys[64]. If, instead, a genome is to be used to study chromosomal evolution, then the VGP-2016 minimum metric 6.7.P5.Q40.C95, with high structural and base accuracies and more than 95% assigned to chromosomes (Table 1), would be necessary. If having GC-rich promoter regions and complete 5′ exons in most genes is essential, then long-read approaches that sequence through these regions are necessary. 'Finished' quality (Table 1) is obviously the ideal assembly result, but this level of quality is currently routine only for bacterial and non-vertebrate model organisms with smaller genome sizes that lack large centromeric satellite arrays[65–67] and for organelle genomes, as presented here[33]. The possibility of achieving complete, telomere-to-telomere assemblies of vertebrate and other eukaryotic species is foreseeable, given some assembled avian and bat chromosomes with zero gaps in this study, and the recent complete assembly of two human chromosomes[68,69].

The Vertebrate Genomes Project

Building on this initial set of assembled genomes and the lessons learned, we propose to expand the VGP to deeper taxonomic phases, beginning with phase 1: representatives of approximately 260 vertebrate orders, defined here as lineages separated by 50 million or more years of divergence from each other. Phase 2 will encompass species that represent all approximately 1,000 vertebrate families; phase 3, all roughly 10,000 genera; and phase 4, nearly all 71,657 extant named vertebrate species (Supplementary Note 5, Supplementary Fig. 3). To accomplish such a project within 10 years, we will need to scale up to completing 125 genomes per week, without sacrificing quality. This includes sample permitting, high molecular weight DNA extractions, sequencing, meta-data tracking, and computational infrastructure. We will take advantage of continuing improvements in genome sequencing technology, assembly, and annotation, including advances in PacBio HiFi reads, Oxford Nanopore reads, and replacements for 10XG reads (Supplementary Note 6), while addressing specific scientific questions at increasing levels of phylogenetic refinement. Genomic technology advances quickly, but we believe the principles of our pipeline and

the lessons learned will be applicable to future efforts. Areas in which improvement is needed include more accurate and complete haplotype phasing, base-call accuracy, and resolution of long repetitive regions such as telomeres, centromeres, and sex chromosomes. The VGP is working towards these goals and making all data, protocols, and pipelines openly available (Supplementary Notes 5, 7).

Despite remaining imperfections, our reference genomes are the most complete and highest quality to date for each species sequenced, to our knowledge. When we began to generate genomes beyond the Anna's hummingbird in 2017, only eight vertebrate species in GenBank had genomes that met our target continuity metrics, and none were haplotype phased (Supplementary Table 23). The VGP pipeline introduced here has now been used to complete assemblies of more than 130 species of similar or higher quality (Supplementary Note 5; BioProject PRJNA489243). We encourage the scientific community to use and evaluate the assemblies and associated raw data, and to provide feedback towards improving all processes for complete and error-free assembled genomes of all species.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03451-0.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Sulston, J. et al. The *C. elegans* genome sequencing project: a beginning. *Nature* **356**, 37–41 (1992).
3. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
4. Howe, K. et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).
5. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* **100**, 659–674 (2009).
6. Koepfli, K.-P., Paten, B., the Genome 10K Community of Scientists & O'Brien, S. J. The Genome 10K Project: a way forward. *Annu. Rev. Anim. Biosci.* **3**, 57–111 (2015).
7. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
8. Adams, M. D. et al. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
9. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
10. Yin, Z.-T. et al. Revisiting avian 'missing' genes from de novo assembled transcripts. *BMC Genomics* **20**, 4 (2019).
11. Korlach, J. et al. 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 (2017).
12. Kelley, D. R. & Salzberg, S. L. Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biol.* **11**, R28 (2010).
13. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
14. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
15. Bradnam, K. R. et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).
16. Zhang, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
17. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
18. Bresler, G., Bresler, M. & Tse, D. Optimal assembly for high throughput shotgun sequencing. *BMC Bioinformatics* **14** (Suppl. 5), S18 (2013).
19. Warren, W. C. et al. The genome of a songbird. *Nature* **464**, 757–762 (2010).
20. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* (2018).
21. Koren, S., Phillippy, A. M., Simpson, J. T., Loman, N. J. & Loose, M. Reply to 'Errors in long-read assemblies can critically affect protein prediction'. *Nat. Biotechnol.* **37**, 127–128 (2019).
22. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
23. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
24. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
25. Howe, K. et al. Significantly improving the quality of genome assemblies through curation. *Gigascience* **10**, giaa153 (2021).
26. Zhou, Y. et al. Platypus and echidna genomes reveal mammalian biology and evolution. *Nature* https://doi.org/10.1038/s41586-020-03039-0 (2021).
27. Kim, J. et al. False gene and chromosome losses affected by assembly and sequence errors. Preprint at https://doi.org/10.1101/2021.04.09.438906 (2021).
28. Lewin, H. A., Graves, J. A. M., Ryder, O. A., Graphodatsky, A. S. & O'Brien, S. J. Precision nomenclature for the new genomics. *Gigascience* **8**, giz086 (2019).
29. Kronenberg, Z. N. et al. Extended haplotype phasing of de novo genome assemblies with FALCON-Phase. *Nat. Commun.* https://doi.org/10.1038/s41467-020-20536-y (2021).
30. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
31. Tomaszkiewicz, M., Medvedev, P. & Makova, K. D. Y and W chromosome assemblies: approaches and discoveries. *Trends Genet.* **33**, 266–282 (2017).
32. Kolesnikov, A. A. & Gerasimov, E. S. Diversity of mitochondrial genome organization. *Biochem. (Mosc.)* **77**, 1424–1435 (2012).
33. Formenti, G. et al. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biol.* (in the press).
34. Harrison, G. L. A. et al. Four new avian mitochondrial genomes help get to basic evolutionary questions in the late cretaceous. *Mol. Biol. Evol.* **21**, 974–983 (2004).
35. Zhao, H. et al. The complete mitochondrial genome of the *Anabas testudineus* (Perciformes, Anabantidae). *Mitochondrial DNA A DNA Mapp. Seq. Anal.* **27**, 1005–1007 (2016).
36. Suzuki, A. et al. How the kinetochore couples microtubule force and centromere stretch to move chromosomes. *Nat. Cell Biol.* **18**, 382–392 (2016).
37. Pfenning, A. R. et al. Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846 (2014).
38. Robinson, R. For mammals, loss of yolk and gain of milk went hand in hand. *PLoS Biol.* **6**, e77 (2008).
39. Brandl, K. et al. Yip1 domain family, member 6 (Yipf6) mutation induces spontaneous intestinal inflammation in mice. *Proc. Natl Acad. Sci. USA* **109**, 12650–12655 (2012).
40. Malmstrøm, M. et al. Evolution of the immune system influences speciation rates in teleost fishes. *Nat. Genet.* **48**, 1204–1210 (2016).
41. Japundžić-Žigon, N., Lozić, M., Šarenac, O. & Murphy, D. Vasopressin & oxytocin in control of the cardiovascular system: an updated review. *Curr. Neuropharmacol.* **18**, 14–33 (2020).
42. Cataldo, I., Azhari, A. & Esposito, G. A review of oxytocin and arginine-vasopressin receptors and their modulation of autism spectrum disorder. *Front. Mol. Neurosci.* **11**, 27 (2018).
43. Warren, W. C. et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175–183 (2008).
44. Ko, B. J. et al. Widespread false gene gains caused by duplication errors in genome assemblies. Preprint at https://doi.org/10.1101/2021.04.09.438957 (2021).
45. Lemaire, S. et al. Characterizing the interplay between gene nucleotide composition bias and splicing. *Genome Biol.* **20**, 259 (2019).
46. Zhang, L., Kasif, S., Cantor, C. R. & Broude, N. E. GC/AT-content spikes as genomic punctuation marks. *Proc. Natl Acad. Sci. USA* **101**, 16855–16860 (2004).
47. Jarvis, E. D. et al. Global view of the functional molecular organization of the avian cerebrum: mirror images and functional columns. *J. Comp. Neurol.* **521**, 3614–3665 (2013).
48. Kubikova, L., Wada, K. & Jarvis, E. D. Dopamine receptors in a songbird brain. *J. Comp. Neurol.* **518**, 741–769 (2010).
49. Sémon, M. & Wolfe, K. H. Rearrangement rate following the whole-genome duplication in teleosts. *Mol. Biol. Evol.* **24**, 860–867 (2007).
50. Jebb, D. et al. Six reference-quality genomes reveal evolution of bat adaptations. *Nature* **583**, 578–584 (2020).
51. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
52. Warren, W. C. et al. A new chicken genome assembly provides insight into avian genome structure. *G3 (Bethesda)* **7**, 109–117 (2017).
53. Meredith, R. W. et al. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524 (2011).
54. Rodriguez-Agudo, D. et al. StarD5: an ER stress protein regulates plasma membrane and intracellular cholesterol homeostasis. *J. Lipid Res.* **60**, 1087–1098 (2019).
55. Kim, J. et al. Reconstruction and evolutionary history of eutherian chromosomes. *Proc. Natl Acad. Sci. USA* **114**, E5379–E5388 (2017).
56. Lin, B., Dutta, B. & Fraser, I. D. C. Systematic investigation of multi-TLR sensing identifies regulators of sustained gene activation in macrophages. *Cell Syst.* **5**, 25–37.e3 (2017).
57. Theofanopoulou, C., Gedman, G. L., Cahill, J. A., Boeckx, C. & Jarvis, E. D. Universal nomenclature for oxytocin-vasotocin ligand and receptor families. *Nature* https://doi.org/10.1038/s41586-020-03040-7 (2021).
58. Ocampo Daza, D. & Haitina, T. Reconstruction of the carbohydrate 6-O sulfotransferase gene family evolution in vertebrates reveals novel member, CHST16, lost in amniotes. *Genome Biol. Evol.* **12**, 993–1012 (2020).
59. Damas, J. et al. Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc. Natl Acad. Sci. USA* **117**, 22311–22322 (2020).
60. Dussex, N. et al. Population genomics reveals the impact of long-term small population size in the critically endangered kākāpō. *Cell Genom.* (in the press).
61. Teeling, E. C. et al. Bat biology, genomes, and the Bat1K project: to generate chromosome-level genomes for all living bat species. *Annu. Rev. Anim. Biosci.* **6**, 23–46 (2018).
62. Lewin, H. A. et al. Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl Acad. Sci. USA* **115**, 4325–4333 (2018).
63. Jarvis, E. D. et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
64. Li, S. et al. Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biol.* **15**, 557 (2014).

65. Koren, S. & Phillippy, A. M. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* **23**, 110–120 (2015).
66. Jenjaroenpun, P. et al. Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of *Saccharomyces cerevisiae* CEN.PK113-7D. *Nucleic Acids Res.* **46**, e38 (2018).
67. Tyson, J. R. et al. MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Res.* **28**, 266–274 (2018).
68. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
69. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* https://doi.org/10.1038/s41586-021-03420-7 (2021).
70. Beçak, M. L., Beçak, W., Roberts, F. L., Shoffner, R. N. & Volpe, P. (eds.) *Chromosome Atlas: Fish, Amphibians, Reptiles, and Birds* Vol. 2 (Springer, 1973).
71. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
72. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

Arang Rhie[1,103], Shane A. McCarthy[2,3,103], Olivier Fedrigo[4,103], Joana Damas[5], Giulio Formenti[4,6], Sergey Koren[1], Marcela Uliano-Silva[7,8], William Chow[3], Arkarachai Fungtammasan[9], Juwan Kim[10], Chul Lee[10], Byung June Ko[11], Mark Chaisson[12], Gregory L. Gedman[6], Lindsey J. Cantin[6], Francoise Thibaud-Nissen[13], Leanne Haggerty[14], Iliana Bista[2,3], Michelle Smith[3], Bettina Haase[4], Jacquelyn Mountcastle[4], Sylke Winkler[15,16], Sadye Paez[4,6], Jason Howard[17], Sonja C. Vernes[18,19,20], Tanya M. Lama[21], Frank Grutzner[22], Wesley C. Warren[23], Christopher N. Balakrishnan[24], Dave Burt[25], Julia M. George[26], Matthew T. Biegler[6], David Iorns[27], Andrew Digby[28], Daryl Eason[28], Bruce Robertson[29], Taylor Edwards[30], Mark Wilkinson[31], George Turner[32], Axel Meyer[33], Andreas F. Kautt[33,34], Paolo Franchini[33], H. William Detrich III[35], Hannes Svardal[36,37], Maximilian Wagner[38], Gavin J. P. Naylor[39], Martin Pippel[15,40], Milan Malinsky[3,41], Mark Mooney[42], Maria Simbirsky[9], Brett T. Hannigan[9], Trevor Pesout[43], Marlys Houck[44], Ann Misuraca[44], Sarah B. Kingan[45], Richard Hall[45], Zev Kronenberg[45], Ivan Sović[45,46], Christopher Dunn[45], Zemin Ning[3], Alex Hastie[47], Joyce Lee[47], Siddarth Selvaraj[48], Richard E. Green[43,49], Nicholas H. Putnam[50], Ivo Gut[51,52], Jay Ghurye[49,53], Erik Garrison[43], Ying Sims[3], Joanna Collins[3], Sarah Pelan[3], James Torrance[3], Alan Tracey[3], Jonathan Wood[3], Robel E. Dagnew[12], Dengfeng Guan[2,54], Sarah E. London[55], David F. Clayton[56], Claudio V. Mello[57], Samantha R. Friedrich[57], Peter V. Lovell[57], Ekaterina Osipova[15,40,58], Farooq O. Al-Ajli[59,60,61], Simona Secomandi[62], Heebal Kim[10,11,63], Constantina Theofanopoulou[6], Michael Hiller[64,65,66], Yang Zhou[67], Robert S. Harris[68], Kateryna D. Makova[68,69,70], Paul Medvedev[69,70,71,72], Jinna Hoffman[13], Patrick Masterson[13], Karen Clark[13], Fergal Martin[14], Kevin Howe[14], Paul Flicek[14], Brian P. Walenz[1], Woori Kwak[63,73], Hiram Clawson[43], Mark Diekhans[43], Luis Nassar[43], Benedict Paten[43], Robert H. S. Kraus[33,74], Andrew J. Crawford[75], M. Thomas P. Gilbert[76,77], Guojie Zhang[78,79,80,81], Byrappa Venkatesh[82], Robert W. Murphy[83], Klaus-Peter Koepfli[84], Beth Shapiro[85,86], Warren E. Johnson[84,87,88], Federica Di Palma[89], Tomas Marques-Bonet[90,91,92,93], Emma C. Teeling[94], Tandy Warnow[95], Jennifer Marshall Graves[96], Oliver A. Ryder[44,97], David Haussler[43,85], Stephen J. O'Brien[98,99], Jonas Korlach[45], Harris A. Lewin[5,100,101], Kerstin Howe[3,104] ✉, Eugene W. Myers[15,40,102,104] ✉, Richard Durbin[2,3,104] ✉, Adam M. Phillippy[1,104] ✉ & Erich D. Jarvis[4,6,86,104] ✉

[1]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [2]Department of Genetics, University of Cambridge, Cambridge, UK. [3]Wellcome Sanger Institute, Cambridge, UK. [4]Vertebrate Genome Lab, The Rockefeller University, New York, NY, USA. [5]The Genome Center, University of California Davis, Davis, CA, USA. [6]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA. [7]Leibniz Institute for Zoo and Wildlife Research, Department of Evolutionary Genetics, Berlin, Germany. [8]Berlin Center for Genomics in Biodiversity Research, Berlin, Germany. [9]DNAnexus Inc., Mountain View, CA, USA. [10]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea. [11]Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea. [12]University of Southern California, Los Angeles, CA, USA. [13]National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD, USA. [14]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. [15]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [16]DRESDEN-concept Genome Center, Dresden, Germany. [17]Novogene, Durham, NC, USA. [18]Neurogenetics of Vocal Communication Group, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. [19]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands. [20]School of Biology, University of St Andrews, St Andrews, UK. [21]University of Massachusetts Cooperative Fish and Wildlife Research Unit, Amherst, MA, USA. [22]School of Biological Science, The Environment Institute, University of Adelaide, Adelaide, South Australia, Australia. [23]Bond Life Sciences Center, University of Missouri, Columbia, MO, USA. [24]Department of Biology, East Carolina University, Greenville, NC, USA. [25]UQ Genomics, University of Queensland, Brisbane, Queensland, Australia. [26]Department of Biological Sciences, Clemson University, Clemson, SC, USA. [27]The Genetic Rescue Foundation, Wellington, New Zealand. [28]Kākāpō Recovery, Department of Conservation, Invercargill, New Zealand. [29]Department of Zoology, University of Otago, Dunedin, New Zealand. [30]University of Arizona Genetics Core, Tucson, AZ, USA. [31]Department of Life Sciences, Natural History Museum, London, UK. [32]School of Natural Sciences, Bangor University, Gwynedd, UK. [33]Department of Biology, University of Konstanz, Konstanz, Germany. [34]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [35]Department of Marine and Environmental Sciences, Northeastern University Marine Science Center, Nahant, MA, USA. [36]Department of Biology, University of Antwerp, Antwerp, Belgium. [37]Naturalis Biodiversity Center, Leiden, The Netherlands. [38]Institute of Biology, Karl-Franzens University of Graz, Graz, Austria. [39]Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. [40]Center for Systems Biology, Dresden, Germany. [41]Zoological Institute, University of Basel, Basel, Switzerland. [42]Tag.bio, San Francisco, CA, USA. [43]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA. [44]San Diego Zoo Global, Escondido, CA, USA. [45]Pacific Biosciences, Menlo Park, CA, USA. [46]Digital BioLogic, Ivanić-Grad, Croatia. [47]Bionano Genomics, San Diego, CA, USA. [48]Arima Genomics, San Diego, CA, USA. [49]Dovetail Genomics, Santa Cruz, CA, USA. [50]Independent Researcher, Santa Cruz, CA, USA. [51]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. [52]Universitat Pompeu Fabra, Barcelona, Spain. [53]Department of Computer Science, University of Maryland College Park, College Park, MD, USA. [54]School of Computer Science and Technology, Center for Bioinformatics, Harbin Institute of Technology, Harbin, China. [55]Department of Psychology, Institute for Mind and Biology, University of Chicago, Chicago, IL, USA. [56]Department of Genetics and Biochemistry, Clemson University, Clemson, SC, USA. [57]Department of Behavioral Neuroscience, Oregon Health and Science University, Portland, OR, USA. [58]Max Planck Institute for the Physics of Complex Systems, Dresden, Germany. [59]Monash University Malaysia Genomics Facility, School of Science, Selangor Darul Ehsan, Malaysia. [60]Tropical Medicine and Biology Multidisciplinary Platform, Monash University Malaysia, Selangor Darul Ehsan, Malaysia. [61]Qatar Falcon Genome Project, Doha, Qatar. [62]Department of Biosciences, University of Milan, Milan, Italy. [63]eGnome, Inc., Seoul, Republic of Korea. [64]LOEWE Centre for Translational Biodiversity Genomics, Frankfurt, Germany. [65]Senckenberg Research Institute, Frankfurt, Germany. [66]Goethe-University, Faculty of Biosciences, Frankfurt, Germany. [67]BGI-Shenzhen, Shenzhen, China. [68]Department of Biology, Pennsylvania State University, University Park, PA, USA. [69]Center for Medical Genomics, Pennsylvania State University, University Park, PA, USA. [70]Center for Computational Biology and Bioinformatics, Pennsylvania State University, University Park, PA, USA. [71]Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA. [72]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA. [73]Hoonygen, Seoul, Korea. [74]Department of Migration, Max Planck Institute of Animal Behavior, Radolfzell, Germany. [75]Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia. [76]Center for Evolutionary Hologenomics, The GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. [77]University Museum, NTNU, Trondheim, Norway. [78]China National Genebank, BGI-Shenzhen, Shenzhen, China. [79]Villum Center for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [80]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [81]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. [82]Institute of Molecular and Cell Biology, A*STAR, Biopolis, Singapore, Singapore. [83]Centre for Biodiversity, Royal Ontario Museum, Toronto, Ontario, Canada. [84]Smithsonian Conservation Biology Institute, Center for Species Survival, National Zoological Park, Washington, DC, USA. [85]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, USA. [86]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [87]The Walter Reed Biosystematics Unit, Museum Support Center MRC-534, Smithsonian Institution, Suitland, MD, USA. [88]Walter Reed Army Institute of Research, Silver Spring, MD, USA. [89]Department of Biological Sciences, Earlham Institute, University of East Anglia, Norwich, UK. [90]Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain. [91]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain. [92]Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [93]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain. [94]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland. [95]Department of Computer Science, The University of Illinois at Urbana-Champaign, Urbana, IL, USA. [96]School of Life Science, La Trobe University, Melbourne, Victoria, Australia. [97]Department of Evolution, Behavior, and Ecology, University of California San Diego, La Jolla, CA, USA. [98]Laboratory of Genomics Diversity-Center for Computer Technologies, ITMO University, St. Petersburg, Russian Federation. [99]Guy Harvey Oceanographic Center, Halmos College of Natural Sciences and Oceanography, Nova Southeastern University, Fort Lauderdale, FL, USA. [100]Department of Evolution and Ecology, University of California Davis, Davis, CA, USA. [101]John Muir Institute for the Environment, University of California Davis, Davis, CA, USA. [102]Faculty of Computer Science, Technical University Dresden, Dresden, Germany. [103]These authors contributed equally: Arang Rhie, Shane A. McCarthy, Olivier Fedrigo. [104]These authors jointly supervised this work: Kerstin Howe, Eugene W. Myers, Richard Durbin, Adam M. Phillippy, Erich D. Jarvis. ✉e-mail: kj2@sanger.ac.uk; gene@mpi-cbg.de; rd109@cam.ac.uk; adam.phillippy@nih.gov; ejarvis@rockefeller.edu

Methods

Genome assembly naming

For each completed assembly of an individual, we gave that assembly an abbreviated name with the following rules: Lineage/GenusSpecies/Individual#.Assembly#. The first letter, in lowercase, identifies the particular lineage: m, mammals; b, birds; r, reptiles; a, amphibians; f, teleost fish; and s, sharks and other cartilaginous fishes. The next three letters (first in caps) identify the species scientific genus name; the next three letters (first in caps) identifies the specific species name. In the last position is the genome identifier, where integers (1, 2, 3, ...) represent different individuals of the same species, and decimals (1.1, 1.2, 1.3, ...) represent different assemblies of the same individual. For example, the first submission of the curated Anna's hummingbird (*Calypte anna*) assembly is bCalAnn1.1, and an updated assembly for the same individual is bCalAnn1.2. When the abbreviated lineage or genus and species names for two or more species were identical, we replaced the subsequent letters (fourth, fifth and so on) of the genus or species name until they could be differentiated. We have created abbreviated names for all 71,657 vertebrate species (http://vgpdb.snu.ac.kr/splist/; https://id.tol.sanger.ac.uk/).

Sample collection

The production of high-quality genome assemblies required us to obtain high-quality cells or tissue that would yield high-molecular-weight (HMW) DNA for long-read sequencing technologies (CLR and ONT) and optical mapping (Bionano). Therefore, we obtained fresh-frozen samples of various tissues (Supplementary Table 8). All samples were obtained according to approved protocols of the respective animal care and use committees or permits obtained by the respective persons and institutions listed in Supplementary Table 8. Additional details of the samples are on their respective BioSample pages (https://www.ncbi.nlm.nih.gov/biosample; accession numbers in Supplementary Table 8). All tissue types tested yielded a sufficient quantity and quality of DNA for sequencing and assembly, but we found that blood worked best for species that have nucleated red blood cells (that is, bird and reptiles), and spleen or cultured cells worked best for mammals, as of to date. Analysis of different tissue types will be presented elsewhere (in preparation).

Isolation of high-molecular-weight DNA

Agarose plug DNA isolation. For tissue, HMW DNA was extracted using the Bionano animal tissue DNA isolation fibrous tissue protocol (cat no. RE-013-10; document number 30071), according to the manufacturer's guidelines. A total of 25–30 mg was fixed in 2% formaldehyde and homogenized using the Qiagen TissueRuptor or manual tissue disruption. For nucleated blood, 27–54 µl was used with an adapted protocol (Bionano, personal communication) of the Bionano Prep Blood and Cell Culture DNA Isolation Kit (cat no. RE-130-10). Lysates were embedded into agarose plugs and treated with Proteinase K and RNase A. Plugs were then purified by drop dialysis with 1× TE. DNA quality was assessed using pulse field gel electrophoresis (PFGE) (Pippin Pulse, SAGE Science, Beverly, MA) or the Femto Pulse instrument (Agilent). PFGE revealed that we isolated ultra-high-molecular-weight DNA between ~100 and ~500 kb long.

Phenol–chloroform gDNA extraction. For some samples, we performed phenol–chloroform extractions for HMW gDNA. Snap-frozen tissue was pulverized into a fine powder with a mortar and pestle in liquid nitrogen. The powdered tissue was lysed overnight at 55 °C in high-salt tissue lysis buffer (400 mM NaCl, 20 mM Tris base (pH 8.0), 30 mM EDTA (pH 8.0), 0.5% SDS, 100 µg/ml Proteinase K), and powdered lung tissue was lysed overnight in Qiagen G2 lysis buffer (cat no. 1014636, Qiagen, Hilden, Germany) containing 100 µg/ml Proteinase K at 55 °C. RNA was removed by incubation in 50 µg/ml RNase A for 1 h at 37 °C. HMW gDNA was purified with two washes of phenol–chloroform-IAA equilibrated to pH 8.0, followed by two washes of chloroform-IAA, and precipitated in ice-cold 100% ethanol. Filamentous HMW gDNA was either spooled with shepherds hooks or collected by centrifugation. HMW gDNA was washed twice with 70% ethanol, dried for 20 min at room temperature and eluted in TE. For the flier cichlid muscle gDNA sample used for PacBio CLR and 10XG libraries, glycogen was precipitated by adding 1/10 (v/v) 0.3 M sodium acetate, pH 6.0 to the extracted genomic DNA, mixing carefully and spinning at room temperature at 10,000*g*. PFGE revealed thatDNA molecule length was between 50 and 300 kb—often lower in size than that obtained with the agarose plug but sufficient for long-range sequencing of CLR and linked read data types.

Others. We also used the Qiagen MagAttract HMW DNA kit (cat no. 67563) and the KingFisher Cell and Tissue DNA kit (Thermo Scientific; cat no. 97030196), following the manufacturers' guidelines. These protocols yielded HMW DNA ranging from 30 to 50 kb. The Genomic Tip (Qiagen) kit was also used for tissue-based extraction of HMW DNA.

Libraries and sequencing

PacBio libraries and sequencing. DNA obtained from agarose plugs was sheared down to ~40 kb fragment size with a MegaRuptor device (Diagenode, Belgium) and fragmented using Covaris g-tubes (520079) or by needle shearing. PacBio large insert libraries were prepared with either the SMRTbell Template Prep Kit 1.0-SPv3 (no.100-991-900) or the SMRTbell Express Template Prep Kit v1 (no. 101-357-000). Libraries were size-selected between 12 and 25 kb using Sage BluePippin (Sage Science, USA), depending on the DNA quality and extraction method. These libraries were sequenced on either RSII or Sequel I instruments, at least 60× coverage per species using Sequel Binding Kit and Sequencing Plate versions 2.0 and 2.1 with 10-h movie time (Supplementary Table 9).

10X Chromium libraries and sequencing. Unfragmented HMW DNA from the agarose plugs was used to generate linked read libraries on 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) following the manufacturer's guidelines. We sequenced the 10X libraries at ~60× coverage per species on an Illumina NovaSeq S4 150-bp PE lane.

Bionano libraries and optical map imaging. Unfragmented ultra-HMW DNA from the agarose plugs was labelled using either two different nicking enzymes (BspQI and BssSI) or a direct labelling enzyme (DLE1) following the Bionano Prep Labelling NLRS (document number 30024) and DLS protocols, respectively (document number 30206). Labelled samples were then imaged on a Bionano Irys or on a Bionano Saphyr instrument. For all species, we aimed for at least 100× coverage per label (Supplementary Table 9).

Hi-C libraries and sequencing. Chromatin interaction (Hi-C) libraries were generated using either Arima Genomics, Dovetail Genomics, or Phase libraries on muscle, blood, or other tissue with in vivo cross-linking (Supplementary Table 9) and sequenced on Illumina instruments. Arima-HiC preparations were performed by Arima Genomics (https://arimagenomics.com/) using the Arima-HiC kit that uses two enzymes (P/N: A510008). The resulting Arima-HiC proximally ligated DNA was then sheared, size-selected around 200–600 bp using SPRI beads, and enriched for biotin-labelled proximity-ligated DNA using streptavidin beads. From these fragments, Illumina-compatible libraries were generated using the KAPA Hyper Prep kit (P/N: KK8504). The resulting libraries were PCR amplified and purified with SPRI beads. The quality of the final libraries was checked with qPCR and Bioanalyzer, and then sequenced on Illumina HiSeq X at ~60× coverage following the manufacturer's protocols. Dovetail-HiC preparations were performed

by Dovetail using a single-enzyme (DpnII) proximity ligation approach. Phase-HiC libraries were made by Phase Genomics using a Proximo Hi-C Library single-enzyme reaction.

Quality control

Before we performed any assembly, all genomic data of all data types from each sample were used to screen potential outlier libraries, outlier sequencing runs, or accidental species contamination with Mash[73] by measuring sequence similarity (Supplementary Fig. 4). When running Mash, we used 21-mers to generate sketches with sketch size of 10,000 and compared among each sequencing run, and then differences assessed between sequencing sets.

Genome size, repeat content, and heterozygosity estimations

These estimations were made with *k*-mer-based methods applied to the Illumina short reads obtained from 10XG linked sequencing libraries. After trimming off barcodes during scaff10x[74] preprocessing, canonical 31-mer counts were collected using Meryl[23]. With the resulting 31-mer histogram, GenomeScope[71] was used to estimate the haploid genome length, repeat content, and heterozygosity. The thorny skate linked read data failed quality control, which we suspect was due to low complexity sequences from the high repeat content (54.1%) of the genome; so *k*-mers were collected later from Illumina whole-genome sequencing reads instead. The genome size and repeat content of the channel bull blenny were estimated from an alternative method that looks at the mode of long read overlap coverage and WindowMasker[75], as the estimated genome size from GenomeScope was almost doubling the known haploid genome size (1.29 Gb versus 0.6 Gb) and repeat content (28.0% versus 58.0%), for reasons related to either the quality of the 10X data or species differences.

Benchmarking assembly steps with the Anna's hummingbird

To develop the VGP standard pipeline, we compared various scaffolding, gap filling, and polishing tools. Default options were used unless otherwise noted. Detailed software versions are listed in Supplementary Table 2.

Contigging and scaffolding. FALCON[76] and FALCON-Unzip[17] (smrtanalysis 3.0.0) were used to generate contigs that used CLR. Canu[77] 1.5+67 was used to generate the combined PacBio CLR and Oxford Nanopore ONT assembly. To benchmark scaffolding with linked reads, we used scaff10x[74] 2.0. For the linked read-only assembly, Supernova 2[78] was used. For the optical maps, two-enzyme hybrid scaffolding was used in the Bionano Solve v3.2.1 software, using BspQI and BssSI initially, as well as DLE1 later when the technology was developed. For benchmarking Hi-C in scaffolding, Salsa 2.2[79] was used for scaffolding results in Fig. 1a, with Hi-C reads generated from Arima Genomics. Additional comparisons for the Hi-C libraries were performed using assemblies provided by Dovetail Genomics and Phase Genomics (Supplementary Table 3). We used Hi-C from Arima Genomics as it had the smallest number of PCR duplicates and better coverage for short and long interactions at the time of comparison (Supplementary Fig. 1). Assembly statistics from HiRise, Proximo HiC, 3D-DNA[80] and Arima Hi-C are available in Supplementary Table 3. We concluded that all Hi-C scaffolding algorithms had similar performance. We decided to use Salsa, as HiRise and Proximo HiC were not open access, and 3D-DNA was computationally expensive on the DNAnexus platform. For short read assemblies, other than Supernova and the NRGene assembly, the assembly GCA_000699085.1[16] was used for benchmarking, which was generated with Illumina paired-end, multiple mate-pair libraries and the SoapDeNovo[81] assembler. The NRGene assembly was provided by the company with DeNovo Magic.

Gap filling. We ran PBJelly with `support --capturedOnly --spanOnly` parameters, to avoid greedy gap closures with no spanning read support. For conservatively filling sequences, we compared different parameters in `output` stage with `--minreads 1` and `--minreads 4` in addition to no restrictions. We found that the number of gaps closed was similar to the gaps filled with Arrow[76] (Supplementary Table 4) and chose not to run PBJelly[82] for future assemblies.

Short-read polishing. Illumina polishing benchmarking was performed using Longranger[83] 2.1.3 and Pilon[84] 1.21 with `--fix bases, local` option (Supplementary Table 5). Later, for the VGP pipeline, we used FreeBayes[85] as Pilon[84] was not computationally scalable for large genomes with the updated Longranger 2.2.2.

Base-level accuracy estimate. Base-level accuracy was measured using a mapping-based approach and later using the *k*-mer-based approach[23]. To determine the number of rounds to polish, we used Illumina paired-end reads from the hummingbird[16].

Mis-joins and missed-joins. The curated hummingbird assembly was mapped to the target assemblies with MashMap2[86] with `--filter_mode one-to-one --pi 95` using 5 kb segments (`-s 5000`) for CLR assemblies and 1 kb (`-s 1000`) for SR assemblies to compensate for the shorter contig sizes, as contigs smaller than a segment size will be excluded from the alignment. The number of mis-joins and missed joins were identified using the assembly_comparison.pl used in the 'Curation' section below (Supplementary Methods, Supplementary Fig. 5).

VGP standard genome assembly pipeline 1.0 to 1.6

All 17 genomes were assembled with the VGP pipeline (Extended Data Fig. 2a) for benchmark purposes, with some uncurated. The pale spear-nosed bat, greater horseshoe bat, Canada lynx, platypus, male and female zebra finch, kākāpō, Anna's hummingbird, Goode's thornscrub tortoise, flier cichlid, and blunt-snouted clingfish assemblies were generated using the VGP pipeline 1.0 to 1.6 and curated for submission to NCBI and EBI public archives. The curated and submitted two-lined caecilian, zig-zag eel, climbing perch, channel bull blenny, eastern happy, and thorny skate assemblies were generated using a similar process developed in parallel (Supplementary Note 2). Two submitted curated versions of the female zebra finch were made, one using the standard VGP pipeline and the other using the VGP trio pipeline, so that comparative analyses could be performed by others.

Contigging. For PacBio data, contigs were generated from subreads using FALCON[76] and FALCON-Unzip[17], with one round of Arrow polishing (smrtanalysis 5.1.0.26412). A minimum read length of 2 kb or a cutoff at which reads longer than the cutoff include 50× coverage was used, whichever was longer. For calculating read coverage, we used estimated genome size from http://www.genomesize.com/ when available, or from the literature (Supplementary Table 11) while waiting for 10XG sequencing to estimate genome size using *k*-mers. FALCON and FALCON-Unzip were run with default parameters, except for computing the overlaps. Raw read overlaps were computed with DALIGNER parameters `-k14 -e0.75 -s100 -l2500 -h240 -w8` to better reflect the higher error rate in early PacBio sequel I and II. Pread (preassembled read) overlaps were computed with DALIGNER parameters `-k24 -e.90 -s100 -l1000 -h600` intending to collapse haplotypes for the FALCON step to better unzip genomes with high heterozygosity rate. FALCON-Unzip outputs both a pseudo-haplotype and a set of alternate haplotigs that represent the secondary alleles. We refer to these outputs as the primary contig set (c1) and alternate contig set (c2).

Purging false duplications. Heterotype false duplications occurred despite setting FALCON[76] parameters to resolve up to 10% haplotype divergence. FALCON-Unzip[17] also incorrectly retained some secondary alleles in the primary contig set, which appeared as false duplications. To reduce these false duplications, we ran Purge_Haplotigs[13], first during

curation (VGP v1.0 pipeline) and then later after contig formation (VGP v1.5 pipeline). To do the former, Purge_Haplotigs was run on the primary contigs (c1), and identified haplotigs were mapped to the scaffolded primary assembly with MashMap2[86] for removal. In the latter, identified haplotigs were moved from the primary contigs (c1) to the alternate haplotig set (p2). The remaining primary contigs were referred to as p1; p2 combined with c2 was referred to as q2. Later, in the VGP v1.6 pipeline, we replaced Purge_Haplotigs with Purge_Dups[14], a new program developed by several of the authors in response to Purge_Haplotigs not removing partial false duplication at contig boundaries. Purging also removes excessive low-coverage (junk) and high-coverage (repeats) contigs. To calculate the presence and overall success of purging false duplications, we used a *k*-mer approach (Supplementary Methods, Supplementary Fig. 6).

Scaffolding with 10XG linked reads. The 10X Genomics linked reads were aligned to the primary contigs (p1), and an adjacency matrix was computed from the barcodes using scaff10x[74] v2.0–2.1. Two rounds of scaffolding were performed. The first round was run with parameters -matrix 2000 -reads 12 -link 10, and the second round with parameters -matrix 2000 -reads 8 -link 10. A gap of 100 bp (represented with 'N's) was inserted between joined contigs. The resulting primary scaffold set was named s1.

Scaffolding with Bionano optical maps. Bionano cmaps were generated using the Bionano Pipeline in non-haplotype assembly mode and used to further scaffold the s1 assembly with Bionano Solve v3.2.1[87]. We began with a one-enzyme nick map (BspQI), followed by a two-enzyme nick map (BspQI and BssSI), and then with a DLE-1 one-enzyme non-nicking approach when the later data type became available (Supplementary Table 9). Scaffold gaps were sized according to the software estimate. The resulting scaffold set was named s2.

Scaffolding with Hi-C reads. Hi-C reads were aligned to the s2 scaffolds using the Arima Genomics mapping pipeline[88]. In brief, both ends of a read pair were mapped independently using BWA-MEM[89] with the parameter -B8, and filtered when mapping quality was <10. Chimeric reads containing a restriction enzyme site were trimmed from the restriction site onward, leaving only the 5′ end. The filtered single-read alignments were then rejoined as paired read alignments. The processed alignments were then used for scaffolding with Salsa2[79], which analyses the normalized frequency of Hi-C interactions between all pairs of contig ends to determine a likely ordering and orientation of each. We used parameters -m yes -i 5 -p yes to allow Salsa2 to break potentially mis-assembled contigs and perform five iterations of scaffolding. After feedback from curation, later versions of Salsa were developed, which more conservatively determine the number of iterations (v2.1) and actively break at mis-assemblies (v2.2), and run for the Canada lynx, Goode's thornscrub tortoise, and two-lined caecilian. The restriction enzyme(s) used to generate each library were specified using parameters -e GATC,GANTC for Arima and -e GATC for Dovetail and Phase Genomics Hi-C data. The resulting Hi-C scaffolded assembly was named s3.

Consensus polishing. To polish bases in both haplotypes with minimal alignment bias, we concatenated the alternate haplotig set (c2 in v1.0 or q2 in v1.5–1.6) to the scaffolded primary set (s3) and the assembled mitochondrial genome (mitoVGP in v1.6). We then performed another round of polishing with Arrow (smrtanalysis 5.1.0.26412) using PacBio CLR reads, aligning with pbalign --minAccuracy=0.75 --minLength=50 --minAnchorSize=12 --maxDivergence=30 –concordant --algorithm=blasr --algorithmOptions=--useQuality --maxHits=1 --hitPolicy=random --seed=1 and consensus polishing with variantCaller --skipUnrecognizedContigs haploid -x 5 -q 20 -X120 –v --algorithm=arrow. While this round of polishing resulted in higher QV for all genomes herein considered, we noticed that it was particularly sensitive to the coverage cutoff parameter (-x). This is because Arrow generates a de novo consensus from the mapped reads without explicitly considering the reference sequence. Later, we found that the second round of Arrow polishing sometimes reduced the QV accuracy for some species. Upon investigation, this issue was traced back to option -x 5, which requires at least 5 reads to call consensus. Such low minimum requirements can lead to uneven polishing in low coverage regions. To avoid this behaviour, we suggest to increase the -x close to the half sequence coverage (for example, 30× when 60× was used for assembly) and check QV before moving forward.

For genomes with a combined assembly size larger than 4 Gb, we used Minimap2[90] with parameters -ax map-pb instead of Blasr[91] to overcome reference index size limitations.

Two more rounds of base-pair polishing were performed with linked reads. The reads were aligned with Longranger align 2.2.2, which incorporates the Lauriat for barcode-aware alignment[83]. From the alignments, homozygous mismatches (variants) were called with FreeBayes[83] v1.2.0 using default options. Consensus was called with bcftools consensus[92] with -i'QUAL>1 && (GT=''AA'' || GT = ''Aa'')' -Hla.

VGP Trio Pipeline v1.0–v1.6. The trio pipeline is similarly designed to the standard pipeline, except for the use of parental data (Extended Data Fig. 3b). When parental genomes are available, the child's CLR reads are binned to maternal and paternal haplotypes, and assembled separately as haplotype-specific contigs (haplotigs) using TrioCanu[20]. In brief, parental specific marker *k*-mers were collected using Meryl[23] from the parental Illumina WGS reads of the parents. These markers were filtered and used to bin the child's CLR read. A haplotype was assigned given the markers observed, normalized by the total markers in each haplotype. The subsequent purging, scaffolding, and polishing steps were similarly updated with the use of Purge_Dups[14] (v1.6). We extended binning to linked reads and Hi-C reads, by excluding read pairs that had any parental-specific marker. The binned Hi-C reads were used to scaffold its haplotype assembly, and polished with the binned linked reads from the observation of haplotype switching using the standard polishing approach. During curation, one of the haplotype assemblies with the higher QV and/or contiguity was chosen as the representative haplotype. The heterogametic sex chromosome from the unchosen haplotype was added to the representative assembly. However, while curating several trios, we found that in regions of low divergence between shared parental homogametic sex chromosomes (that is, X or Z), a small fraction of offspring CLR data was mis-assigned to the wrong haplotype. This mis-alignment resulted in a duplicate, low-coverage offspring X or Z assembly in the paternal (for mammals) or maternal (for birds) haplotype, respectively, which required removal during curation. We are working on methods to improve the binning accuracy for resolution of this issue going forward.

For the female zebra finch in particular, contigs were generated before the binning was automated in the Canu assembler as TrioCanu1.7, and therefore a manual binning process was applied as described in the original Trio-binning paper[20] (Supplementary Methods). Contigs were assembled for each haplotype using the binned reads, excluding unclassified reads. The contigs were polished with two rounds of Arrow polishing using the binned reads, and scaffolded following the v1.0 pipeline with no purging. Additional scaffolding rounds with Bionano (s4) and Hi-C were applied. Scaffolds were renamed according to the primary scaffold assembly of the same individual (s5), with sex chromosomes grouped as Z in the paternal assembly and W in the maternal assembly following synteny to the Z chromosome from the curated male zebra finch VGP assembly. Two rounds of SR polishing were applied using linked reads, by mapping on both haplotypes. After haplotype switches were discovered, additional rounds of polishing were applied using binned linked reads (Supplementary Methods).

Mitochondrial genome assembly. Similar to other recent methods[93,94], we developed a reference-guided MT assembly pipeline. MT reads in the raw CLR data were identified by mapping the whole read set to an existing reference sequence of the specific species or of closely related species using Blasr. Filtered mtDNA CLRs were assembled into a single contig using Canu v1.8, polished with Arrow using CLR and then FreeBayes v1.0.2 together with bcftools v1.9 using short reads from the 10XG data (Extended Data Fig. 3c). The overlapping sequences at the ends of the contig were trimmed, and the remaining contig sequence circularized. The mitoVGP pipeline is made available at https://github.com/VGP/vgp-assembly/tree/master/mitoVGP. A more detailed protocol description of the assembly pipeline and new discoveries from the MT assemblies are published elsewhere[33].

Curation

The VGP genome assembly pipeline produces high quality assemblies, yet no automated method to date is free from the production of errors, especially during the scaffolding stages. To minimize the impact of the remaining algorithmic shortcomings, we subjected all assemblies to rigorous manual curation. All data generated for a species in this study and other publicly available data (for example, genetic maps, gene sets and genome assemblies of the same or closely related species) were collated, aligned to the primary assembly and analysed in gEVAL[95] (https://vgp-geval.sanger.ac.uk/index.html), visualizing discordances in a feature browser and issue lists. In parallel, Hi-C data were mapped to the primary assembly and visualized using Juicebox[96] and/or HiGlass[97]. With these data, genome curators identified mis-joins, missed joins and other anomalies, and corrected the primary assembly accordingly. No change was made without unambiguous evidence from available data types; for example, a Hi-C suggested join would not be made unless supported by BioNano maps, long-read data, or gene alignments. When sequencing the heterogametic sex, we identified sex chromosomes based on half coverage, homology alignments to sex chromosomes in other species, and the presence of sex chromosome-specific genes.

Contamination removal. A succession of searches was used to identify potential contaminants in the generated assemblies.

1) A megaBLAST[98] search against a database of common contaminants (ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_euks.fa.gz) requiring $e \leq 1 \times 10^{-4}$, reporting matches with ≥98% sequence identity and match length 50–99 bp, ≥94% and match length 100–199 bp, or ≥90% and match length 200 bp or above.

2) A vecscreen (https://www.ncbi.nlm.nih.gov/tools/vecscreen/) search against a database of adaptor sequences (ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/adaptors_for_screening_euks.fa)

3) After soft-masking repeats using Windowmasker[75], a megaBLAST search against chromosome-level assemblies from RefSeq requiring $e \leq 1 \times 10^{-4}$, match score ≥100, and sequence identity ≥98%; regions matching highly conserved rDNAs were ignored.

Manual inspection of the results was necessary to differentiate contamination from conservation and/or horizontal gene transfer. Adaptor sequences were masked; other contaminant sequences were removed. Assemblies were also checked for runs of Ns at the ends of scaffolds, created as artefacts of the iterative scaffolding process, and when found they were trimmed.

Organelle genomes. These were detected by a megaBLAST search against a database of known organelle genomes requiring $e \leq 1 \times 10^{-4}$, sequence identity ≥90%, and match length ≥500; the databases are available at ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/mito.nt.gz and ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plastid/*genomic.fna.gz. Only scaffolds consisting entirely of organelle sequences were assumed to be organelle genomes, and replaced by the genome from the separate organelle assembly pipeline. Organelle matches embedded in nuclear sequences that were found to be NuMTs were kept.

False duplication removal. Retained false duplications were identified using Purge_Haplotigs[13] run either after scaffolding and polishing (Anna's hummingbird, kākāpō, male zebra finch, female zebra finch, platypus, pale spear-nosed bat, and greater horseshoe bat) or on the c1 before scaffolding (two-lined caecilian, flier cichlid, Canada lynx, and Goode's thornscrub tortoise). Subsequent manual curation identified additional haplotypic duplications for the listed assemblies and also those that were not treated with Purge_Haplotigs (Eastern happy, climbing perch, zig-zag eel). The evidence used included read coverage, sequence self-comparison, transcript alignments, Bionano map alignments and Hi-C 2D maps, all confirming the superfluous nature of one allele. The identified haplotype duplications were moved from the primary to the alternate assembly.

Chromosome assignment. For a scaffold to be annotated as a chromosome, we used evidence from Hi-C as well as genetic linkage or FISH karyotype mapping when available. For Hi-C evidence, we considered a scaffold as a complete chromosome (albeit with gaps) when there was a clear unbroken diagonal in the Juicebox or HiGlass plots for that scaffold and no other large scaffolds that could be joined to that same scaffold; if present and no unambiguous join was possible, we named it as an unlocalized scaffold for that chromosome. When we could not find evidence of a complete chromosome, we kept the scaffold number for its name. We named all evidence-validated scaffolds as chromosomes down to the smallest Hi-C box unit resolution allowed with these characteristics. When there was an established chromosome terminology for a given species or set of species, we use the established terminology except when our new assemblies revealed errors in the older assembly, such as scaffold/chromosome fusions, fissions, rearrangements, and non-chromosome names. For species without an established chromosome terminology, we named the scaffolds as chromosomes numbers 1, 2, 3 …, in descending order of scaffold size. For the sex chromosomes, we used the letters X and Y for mammals and Z and W for birds.

Using comparative genomics to assess assembly structure. In cases where a high-quality chromosome-level genome was available for a closely related species, comparative genome analysis was performed. The polished primary assembly (t3.p) was mapped to the related genome using MashMap2[86] with --pi 75 -s 300000. The number of chromosomal differences was identified using a custom script available at https://github.com/jdamas13/assembly_comparison. This resulted in the identification of ~60 to ~450 regions for each genome assembly flanking putative misassemblies or lineage-specific genome rearrangements. To identify which were real misassemblies, the identified discrepancies were communicated to the curation team for manual verification (see above).

To identify any possible remaining mis-joins, each curated avian and mammalian assembly was compared with the zebra finch (taeGut2) or human (hg38) genomes, respectively. Pairwise alignments between each of the VGP assemblies and the clade reference were generated with LastZ[99] (version 1.04) using the following parameters: *C* = 0 *E* = 30 *H* = 2000 *K* = 3000 *L* = 2200 *O* = 400. The pairwise alignments were converted into the UCSC 'chain' and 'net' formats with axtChain (parameters: -minScore = 1000 -verbose = 0 -linearGap = medium) followed by chainAntiRepeat, chainSort, chainPreNet, chainNet and netSyntenic, all with default parameters[100]. Pairwise synteny blocks were defined using maf2synteny[101] at 100-, 300-, and 500-kb resolutions. Evolutionary breakpoint regions were detected and classified using an ad hoc statistical approach[102]. This analysis identified 2 to 90 genomic regions per assembly that could be flanking misassemblies, lineage-specific chromosome rearrangements, or reference-specific chromosome rearrangements (116 in the human

and 26 in the zebra finch). Determining the underlying cause for each of the flagged regions will need further verification. All alignments are available for visualization at the Evolution Highway comparative chromosome browser (http://eh-demo.ncsa.illinois.edu/vgp/).

Annotation

NCBI and Ensembl annotation pipeline used in this study are described in the Supplementary Methods.

Evaluation

Detailed methods for other types of evaluation, including BUSCO runs, mis-join and missed-join identification, reliable blocks, collapsed repeats, telomeres, RNA-seq and ATAC–seq mapping, and false gene duplications are in the Supplementary Methods. No statistical methods were used to predetermine sample size, the experiments were not randomized, and the investigators were not blinded to group during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All raw data, intermediate and final assemblies are publicly available via GenomeArk (https://vgp.github.io/genomeark), archived on NCBI/EBI BioProject under accession PRJNA489243 with annotations, and browsable on the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/hubs/VGP/). The final primary assembly from the automated pipeline before curation is browsable on gEVAL (https://vgp-geval.sanger.ac.uk) with all four raw data mappings. The VGP assembly pipeline is available as a stand-alone pipeline (https://github.com/VGP/vgp-assembly) as well as a workflow on DNAnexus (https://platform.dnanexus.com/). A VGP-specific assembly hub portal in the U.C. Santa Cruz browser is available as a gateway to access all VGP genome assemblies and annotations (https://hgdownload.soe.ucsc.edu/hubs/VGP).

Code availability

All codes used in the VGP Assembly Pipeline and the VGP Trio Pipeline are publicly available at https://github.com/VGP/vgp-assembly/tree/master/pipeline.

73. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. **17**, 132 (2016).
74. Ning, Z. & Harry, E. Scaff10X https://github.com/wtsi-hpag/Scaff10X.
75. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
76. Chin, C.-S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
77. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. **27**, 722–736 (2017).
78. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res*. **27**, 757–767 (2017).
79. Ghurye, J. et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol*. **15**, e1007273 (2019).
80. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
81. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
82. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
83. Bishara, A. et al. Read clouds uncover variation in complex regions of the human genome. *Genome Res*. **25**, 1570–1580 (2015).
84. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
85. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at http://arxiv.org/abs/1207.3907 (2012).
86. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748–i756 (2018).
87. Bionano Genomics, Inc. Bionano Software Downloads. https://bionanogenomics.com/support/software-downloads/.
88. Arima Genomics, Inc. Arima Genomics Mapping Pipeline. https://github.com/ArimaGenomics/mapping_pipeline.
89. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
90. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
91. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
92. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
93. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res*. **45**, e18 (2017).
94. Soorni, A., Haak, D., Zaitlin, D. & Bombarely, A. Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC Genomics* **18**, 49 (2017).
95. Chow, W. et al. gEVAL—a web-based browser for evaluating genome assemblies. *Bioinformatics* **32**, 2508–2510 (2016).
96. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst*. **3**, 99–101 (2016).
97. Kerpedjiev, P. et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol*. **19**, 125 (2018).
98. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
99. Harris, R. S. Improved Pairwise Alignment of Genomic DNA. Thesis, Pennsylvania State Univ. (2007).
100. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100**, 11484–11489 (2003).
101. Kolmogorov, M., Raney, B., Paten, B. & Pham, S. Ragout—a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* **30**, i302–i309 (2014).
102. Farré, M. et al. Novel insights into chromosome evolution in birds, archosaurs, and reptiles. *Genome Biol. Evol*. **8**, 2442–2451 (2016).
103. Guan, D. Asset. https://github.com/dfguan/asset.
104. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4.10.1–4.10.14 (2009).
105. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
106. Harry, E. PretextView. https://github.com/wtsi-hpag/PretextView.
107. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol*. **5**, R12 (2004).
108. Nattestad, M. Dot. https://github.com/MariaNattestad/dot.

Acknowledgements We thank the following persons for feedback and support: R. Johnson, E. Karlsson, K. Lindblad Toh, W. Jun, I. Korf, W. Haerty, G. Etherington, B. Clavijo, and A. Komissarov for discussions in the early stages of the project; R. Fuller for help with the G10K website maintenance, and H. Segal for help with with VGP website development; M. Linh Pham for help with initial grant writing; L. Shalmiyev for administrative help; D. Church, G. Kol, K. Baruch, O. Barad, I. Liachko, E. Muzychenko, S. Garg, and M. Kolmogorov for preliminary analyses performed on one or more genomes; K. Oliver, C. Corton and J. Skelton for data generation; E. Harry for technical support in scaff10x and Pretext; C. Mazzoni for coordinating students and training at Leibniz Institute for Zoo and Wildlife Research and Berlin Center for Genomics in Biodiversity Research; and M. Driller, C. Caswara, M. Vafadar, N. Hill, D. De Panis, A. Whibley, B. Maloney, C. Mitchell, G. Gallo, J. Gaige, K. Amoako-Boadu, M. Jose Gomez, M. Montero, D. Ratnikov, S. Brown, S. Zylka, S. Marcus, and T. Carrasco for completing training and testing the VGP pipeline by producing ordinal representative genome assemblies not described in this manuscript. We thank our company partners (listed below), NCBI, EBI, and Amazon AWS, including AWS for sponsoring sequence storage. J. Fekecs and D. Leja created the animal images, and J. Kim modified them to silhouettes. We thank them for their permission to publish. A.R., S.K., B.P.W. and A.M.P. were supported by the Intramural Research Program of the NHGRI, NIH (1ZIAHG200398). A.R. was also supported by the Korea Health Technology R&D Project through KHIDI, funded by the Ministry of Health & Welfare, Republic of Korea (HI17C2098). S.A.M., I.B. and R.D. were supported by Wellcome Trust grant WT207492; W.C., M. Smith, Z.N., Y.S., J.C., S. Pelan, J.T., A.T., J.W. and Kerstin Howe by WT206194; L.H., F.M., Kevin Howe and P. Flicek by WT108749/Z/15/Z, WT218328/B/19/Z and the European Molecular Biology Laboratory. O.F. and E.D.J. were supported by Howard Hughes Medical Institute and Rockefeller University start-up funds for this project. J.D. and H.A.L. were supported by the Robert and Rosabel Osborne Endowment. M.U.-S. received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (750747). F.T.-N., J. Hoffman, P. Masterson and K.C. were supported by the Intramural Research Program of the NLM, NIH. C.L., B.J.K., J. Kim and H.K. were supported by the Marine Biotechnology Program of KIMST, funded by the Ministry of Ocean and Fisheries, Republic of Korea (20180430). M.C. was supported by Sloan Research Fellowship (FG-2020-12932). S.C.V. was funded by a Max Planck Research Group award from the Max Planck Society, and a Human Frontiers Science Program (HFSP) Research grant (RGP0058/2016). T.M.L., W.E.J. and the Canada lynx genome were funded by the Maine Department of Inland Fisheries & Wildlife (F11AF01099), including when W.E.J. held a National Research Council Research Associateship Award at the Walter Reed Army Institute of Research (WRAIR). C.B. was supported by the NSF (1457541 and 1456612). D.B. was funded by The University of Queensland (HFSP - RGP0030/2015). D.I. was supported by Science Exchange Inc. (Palo Alto, CA). H.W.D. was supported by NSF grants (OPP-0132032 ICEFISH 2004 Cruise, PLR-1444167 and OPP-1955368) and the Marine Science Center at Northeastern University (416). G.J.P.N. and the thorny skate genome were funded by Lenfest Ocean Program (30884). M.P. was funded by the German Federal Ministry of Education and Research (01IS18026C). M. Malinsky was supported by an EMBO fellowship (ALTF 456-2016). The following authors' contributions were supported by the NIH: S. Selvaraj (R44HG008118); C.V.M., S.R.F., P.V.L. (R21 DC014432/DC/NIDCD); K.D.M.

(R01GM130691); H.C. (5U41HG002371-19); M.D. (U41HG007234); and B.P. (R01HG010485). D.G. was supported by the National Key Research and Development Program of China (2017YFC1201201, 2018YFC0910504 and 2017YFC0907503). F.O.A. was supported by Al-Gannas Qatari Society and The Cultural Village Foundation-Katara, Doha, State of Qatar and Monash University Malaysia. C.T. was supported by The Rockefeller University. M. Hiller was supported by the LOEWE-Centre for Translational Biodiversity Genomics (TBG) funded by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK). H.C. was supported by the NHGRI (5U41HG002371-19). R.H.S.K. was funded by the Max Planck Society with computational resources at the bwUniCluster and BinAC funded by the Ministry of Science, Research and the Arts Baden-Württemberg and the Universities of the State of Baden-Württemberg, Germany (bwHPC-C5). B.V. was supported by the Biomedical Research Council of A*STAR, Singapore. T.M.-B. was funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (864203), MINECO/FEDER, UE (BFU2017-86471-P), Unidad de Excelencia María de Maeztu, AEI (CEX2018-000792-M), a Howard Hughes International Early Career award, Obra Social "La Caixa" and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). E.C.T. was supported by the European Research Council (ERC-2012-StG311000) and an Irish Research Council Laureate Award. M.T.P.G. was supported by an ERC Consolidator Award 681396-Extinction Genomics, and a Danish National Research Foundation Center Grant (DNRF143). T.W. was supported by the NSF (1458652). J. M. Graves was supported by the Australian Research Council (CEO561477). E.W.M. was partially supported by the German Federal Ministry of Education and Research (01IS18026C). Complementary sequencing support for the Anna's hummingbird and several genomes was provided by Pacific Biosciences, Bionano Genomics, Dovetail Genomics, Arima Genomics, Phase Genomics, 10X Genomics, NRGene, Oxford Nanopore Technologies, Illumina, and DNAnexus. All other sequencing and assembly were conducted at the Rockefeller University, Sanger Institute, and Max Planck Institute Dresden genome labs. Part of this work used the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov). We acknowledge funding from the Wellcome Trust (108749/Z/15/Z) and the European Molecular Biology Laboratory. We thank Le Comité Scientifique Régional du Patrimoine Naturel and Direction de l'Environnement, de l'Aménagement et du Logement, Guyanne for research approvals and export permits.

Author contributions Wrote the paper and co-coordinated the study: A.R., E.D.J., A.M.P., R.D., E.W.M., Kerstin Howe, S.A.M., O.F. Coordination with vendors: J. Korlach, S. Selvaraj, R.E.G., A.H., M. Mooney. Collected samples: M.T.P.G., W.E.J., R.W.M., G.Z., B.V., M.T.B., J. Howard, S.C.V., T.M.L., F.G., W.C.W., D.B., J. M. George, M.T.B., D.I., A.D., D.E., B.R., T.E., M. Wilkinson, G.T., A. Meyer, A.F.K., P. Franchini, H.W.D., H.S., M. Wagner, G.J.P.N., R.D., E.D.J., E.C.T., R.H.S.K. Generated genome data: O.F., I.B., M. Smith, B.H., J.M., S.W., C.B., A. Meyer, A.F.K., P. Franchini, I.G., D.F.C., C.V.M. Generated genome assemblies: A.R., S.A.M., S.K., M.P., S.B.K., R.H., J.G., Z.N., J.L., B.P.W., M. Malinsky. Generated/modified software: S.K., A.R., S.B.K., R.H., Z.K., J. Korlach, I.S., C.D., Z.N., A.H., J.L., J.G., E.G., C.V.M., S.R.F., N.H.P. Pipeline development: A.R., S.A.M., G.F., S.K., M.U.-S., A.F., M. Simbirsky, B.T.H., T.P., M.P., E.W.M., R.D., A.M.P. Generated MT assemblies: G.F., J. Korlach. Curation: Kerstin Howe, W.C., Y.S., J.C., S. Pelan, J.T., A.T., J.W., Y.Z., J.D., H.A.L. Sex chromosomes: Y.Z., R.S.H., K.D.M., P. Medvedev, J. M. Graves. Hummingbird karyotype analyses: M. Houck, A. Misuraca, M.P., E.W.M., E.D.J. Annotation: F.T.-N., L.H., J. Hoffman, P. Masterson, K.C., F.M., Kevin Howe, P. Flicek, D.B. Evaluation analysis: A.R., J.D., M.U.-S., J. Kim, C.L., B.J.K., M.C., G.L.G., L.J.C., F.T.-N., L.H., J. M. George, J.G., R.E.D., D.G., S.E.L., D.F.C., C.V.M., S.R.F., P.V.L., E.O., F.O.A.-A., S. Secomandi, C.T., M. Hiller, H.K., Kerstin Howe, E.W.M., R.D., A.M.P., E.D.J. Biological findings: J.D., J. Kim, C.L., B.J.K., G.L.G., L.J.C., H.A.L., A.R., E.D.J. Data availability: A.R., S.A.M., W.C., A.F., S. Paez, M. Simbirsky, B.T.H., B.P.W., W.K., H.C., M.D., L.N., B.P., A.M.P., E.D.J. G10K council, founders, and coordination of VGP: T.M.-B., A.J.C., F.D.P., R.D., M.T.P.G., E.D.J., K.-P.K., H.A.L., R.W.M., E.W.M., E.C.T., B.V., G.Z., A.M.P., S. Paez, J. M. Graves, O.A.R., D.H., S.J.O., T.W. and B.S. All authors reviewed the manuscript.

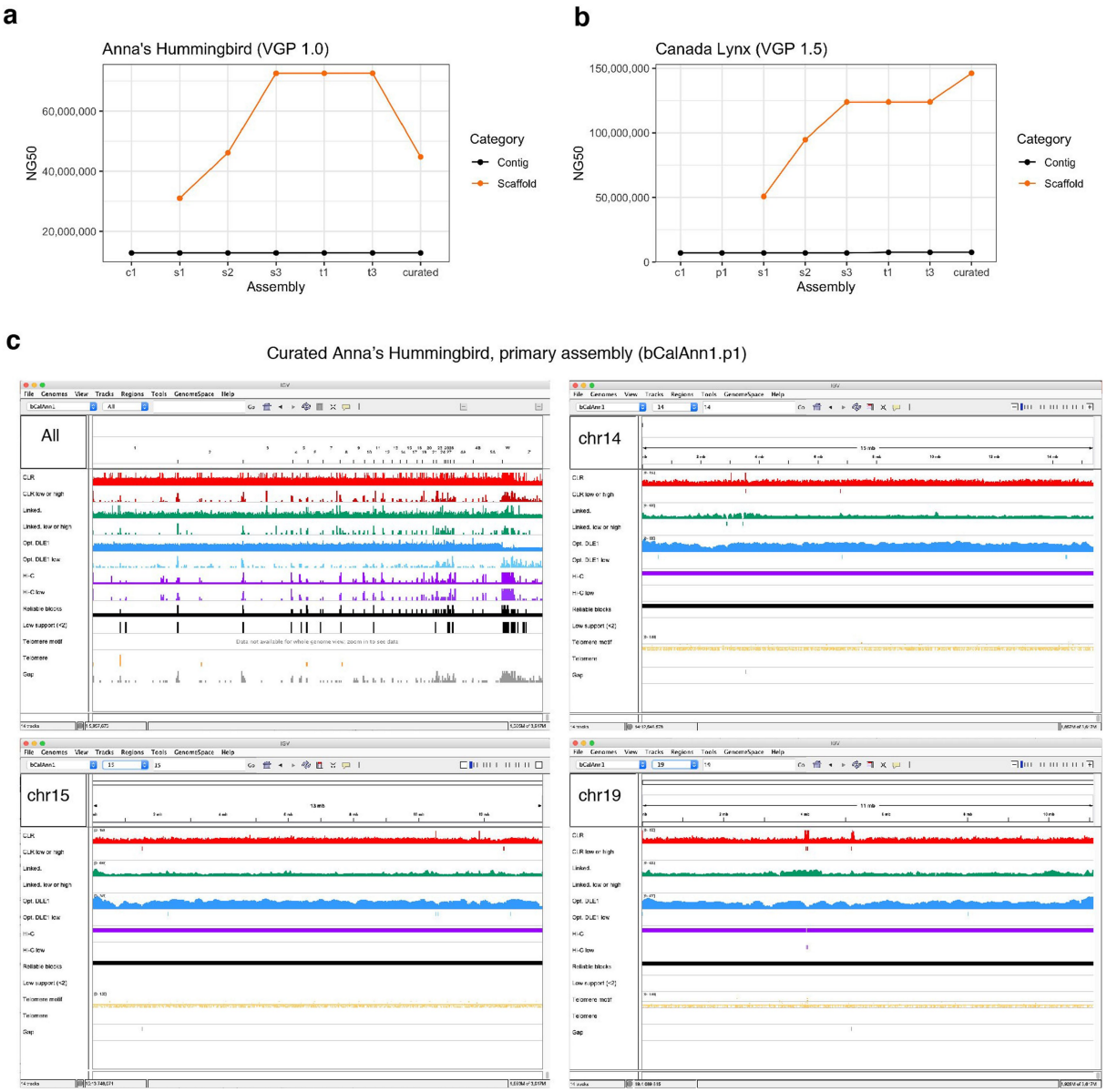
Competing interests During the contributing period, B.T.H., M. Simbirsky, A.F. and M. Mooney were employees of DNAnexus Inc. S.B.K., R.H., Z.K., J. Korlach, I.S. and C.D. were full-time employees at Pacific Biosciences, a company developing single-molecule long read sequencing technologies. R.E.G., N.H.P., and J.G. were affiliated with Dovetail Genomics, a company developing genome assembly tools, including Hi-C. I.G. was affiliated with Oxford Nanopore Technologies, a company generating long read sequencing technologies. A.H. and J.L were employees of Bionano Genomics, a company developing optical maps for genome assembly. S. Selvaraj was an employee of Arima Genomics, a company developing Hi-C data for genome assemblies. R.D. is a scientific advisory board member of Dovetail Inc. P. Flicek is a member of the Scientific Advisory Boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd. H.C. receives royalties from the sale of UCSC Genome Browser source code, LiftOver, GBiB, and GBiC licenses to commercial entities. S.K. has received travel funds to speak at symposia organized by Oxford Nanopore. M.D. and L.N. receive royalties from licensing of UCSC Genome Browser. For W.E.J., the content here is not to be construed as the views of the DA or DOD. All other authors declare no competing interests.

Additional information

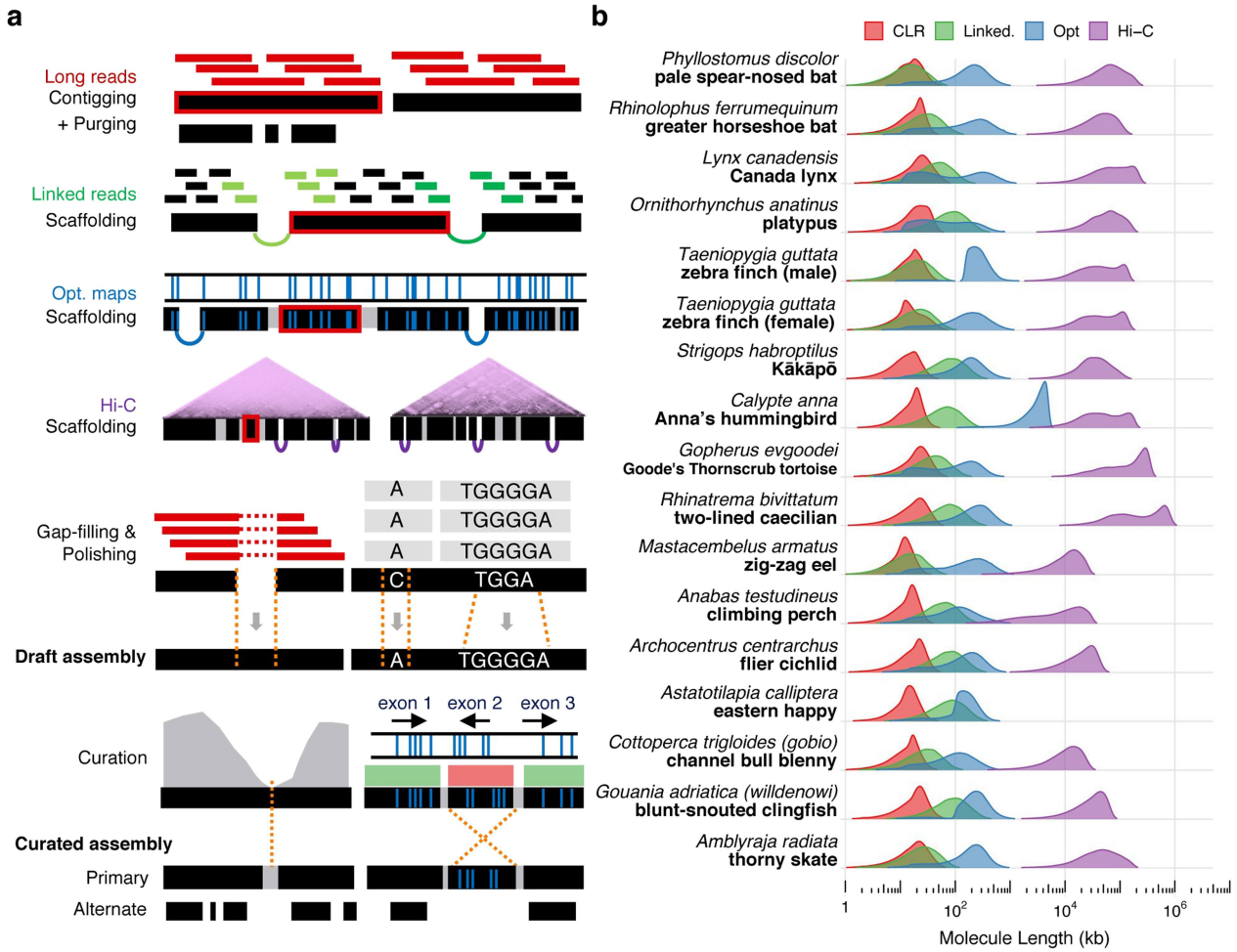
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03451-0.

Correspondence and requests for materials should be addressed to Kerstin Howe., E.W.M., R.D., A.M.P. or E.D.J.

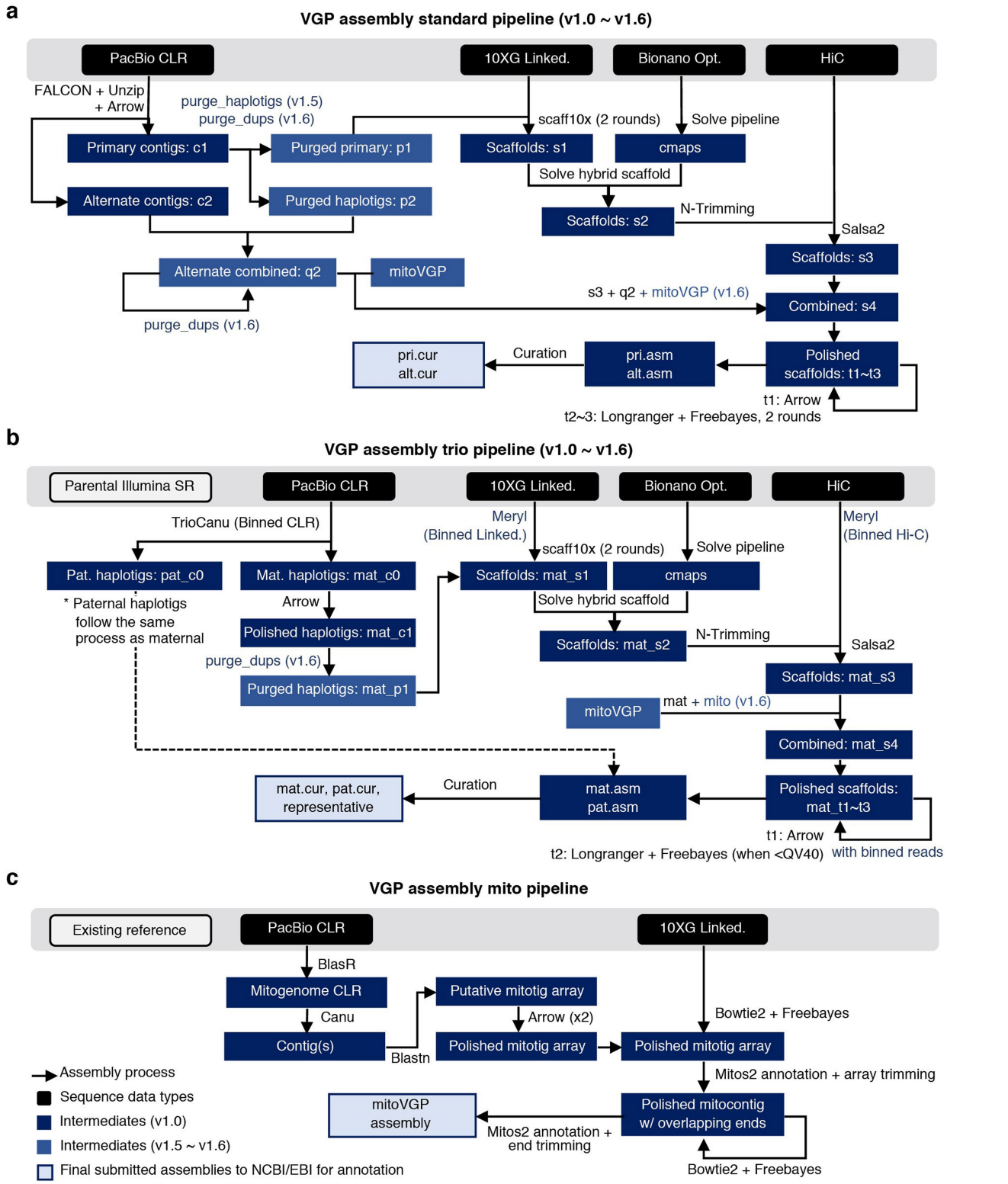
Peer review information *Nature* thanks Michael Schatz, Justin Zook and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



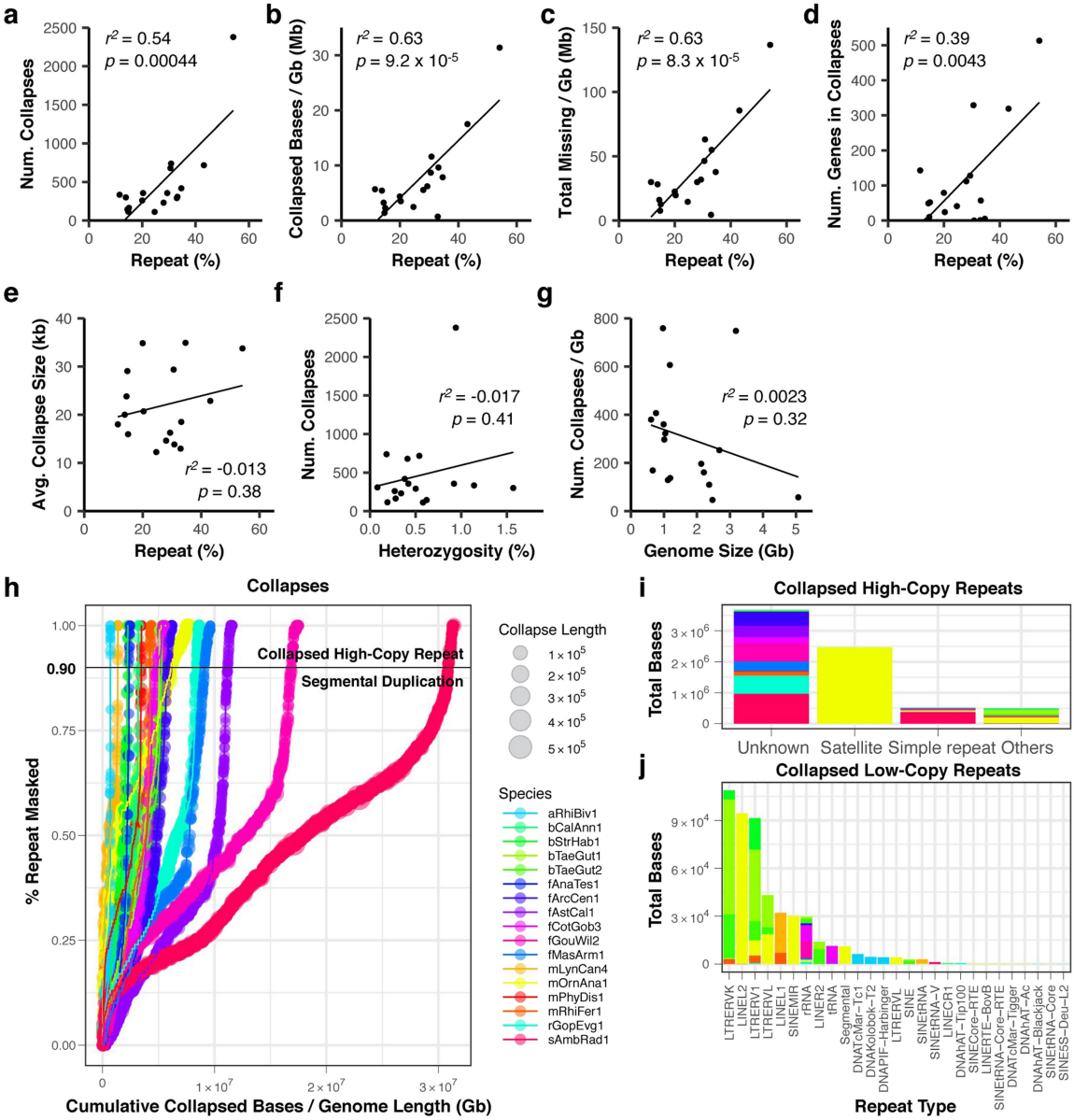
Extended Data Fig. 1 | Assessment of completeness of the Anna's hummingbird assembly. a, **b**, Steps and NG50 continuity values of the VGP assembly pipeline that gave the highest quality assembly for Anna's hummingbird (**a**) and Canada lynx (**b**) in this study. The specific steps are outlined further in Extended Data Fig. 2a, and Methods. **c**, Whole-genome alignment of CLR (red), linked reads (green), optical maps (blue), and Hi-C reads (purple) of the Anna's hummingbird, along with telomere motif (TTAGGG and its reverse complement, yellow) and gaps (grey) using Asset software[103]. For each data type, the first row shows the mapped coverage, and the second shows the number of counts of low coverage or signs of collapsed repeats. Larger chromosomal scaffolds (1–19) have fewer gaps and low coverage or collapsed regions compared with the micro chromosomes (20–33). Chromosomes 14, 15 and 19 of the Anna's hummingbird were the most structurally reliable scaffolds, having only one gap each with no low-support regions. We defined reliable blocks as those supported by at least two technologies. Reliable blocks excluded regions with structural assembly errors, such as collapsed repeats or unresolved segmental duplications. Low-support regions are those where the reliable blocks row has a peak.



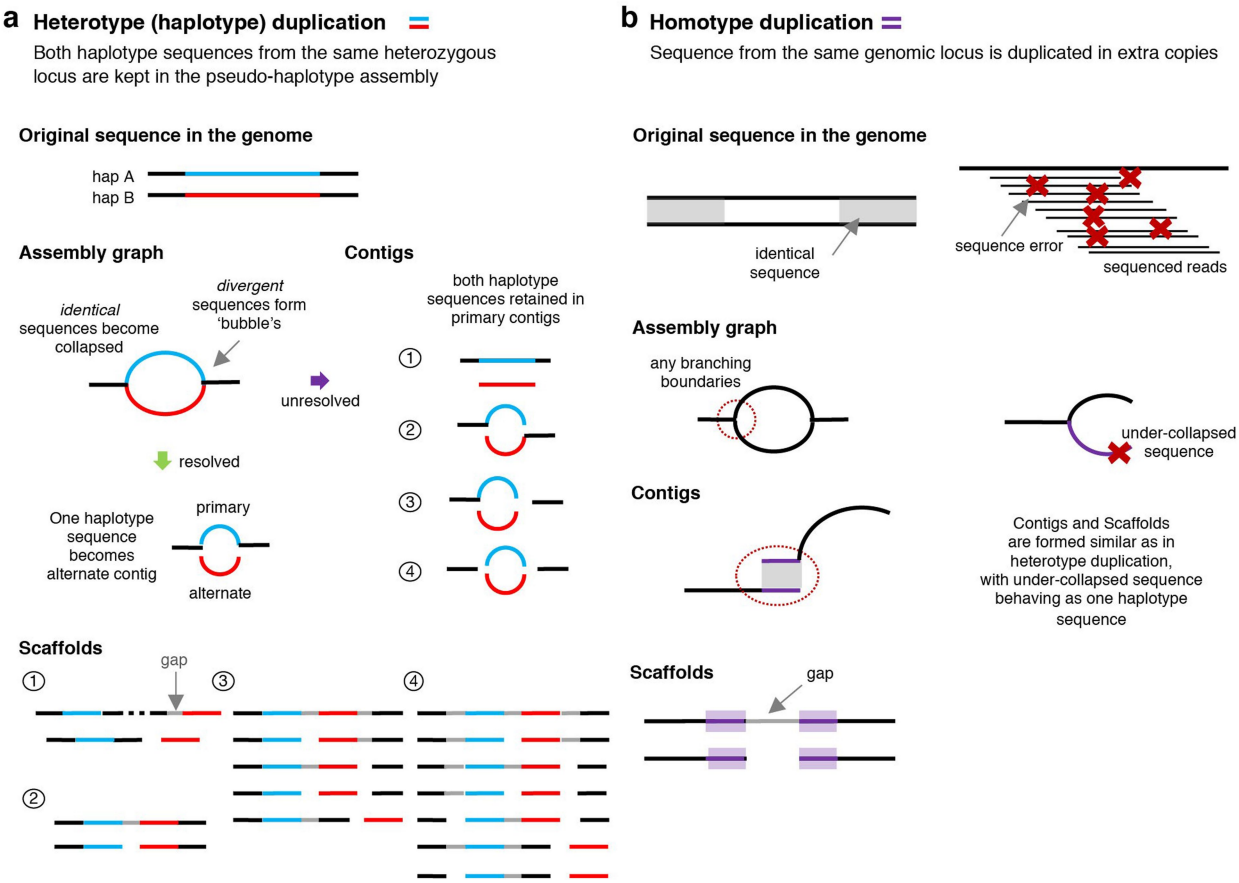
Extended Data Fig. 2 | VGP assembly pipeline applied across multiple species. a, Iterative assembly pipeline of sequence data types (coloured as in **b**) with increasing chromosomal distance. Thin bars, sequence reads; thick black bars, assembled contigs; black bars with space and arcing links, scaffolds; grey bars, gaps placed by previous steps; thick red border, tracking of an example contig in the pipeline. The curation step shows an example of a mis-assembly break identified by sequence coverage (grey, left) and an example of an inversion error (right) detected by the optical map. **b**, Intra-molecule length distribution of the four data types used to generate the assemblies of 16 vertebrate species, weighted by the fraction of bases in each length bin (log scaled). Molecule length above 1 kb was measured from read length for CLR, estimated molecule coverage for linked reads, raw molecule length for optical maps, and interaction distance for Hi-C reads. For each species, the fragment length distribution of each data type was similar to those for the Anna's hummingbird, with differences primarily influenced by tissue type, preservation method, and collection or storage conditions (unpublished data).



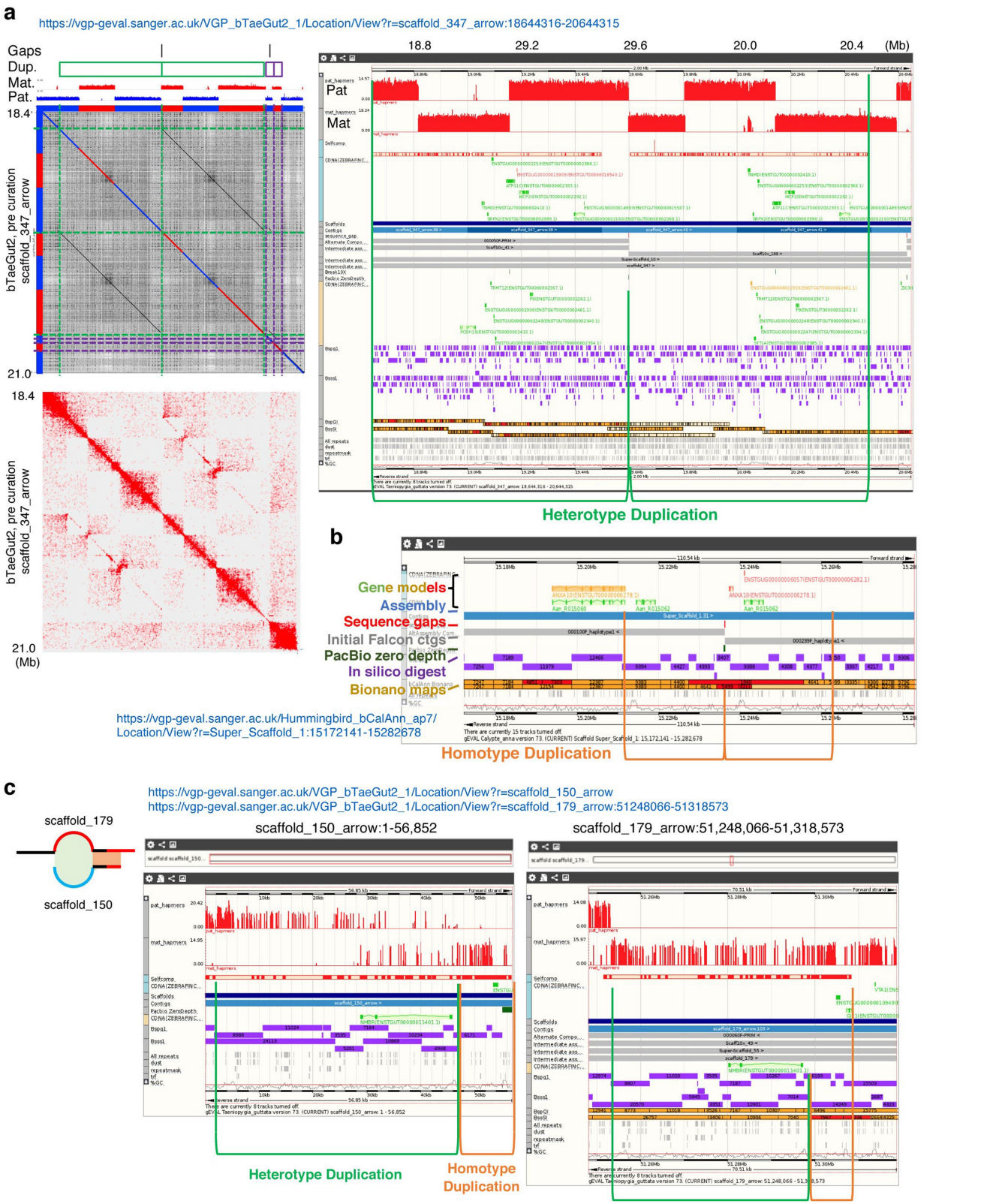
Extended Data Fig. 3 | Flow charts of assembly pipelines used to generate high-quality assemblies in this study. a, Standard VGP assembly pipeline when sequencing data of one individual, that generated the highest quality assemblies: generate primary pseudo-haplotype and alternate haplotype contigs with CLR using FALCON-Unzip[17]; generate scaffolds with linked reads using Scaff10x[74]; break mis-joins and further scaffold with optical maps using Solve[87]; generate chromosome-scale scaffolds with Hi-C reads using Salsa2[79]; fill in gaps and polish base-errors with CLR using Arrow (Pacific BioSciences); perform two or more rounds of short-read polishing with linked reads using FreeBayes[85]; and perform expert manual curation to correct potential assembly errors using gEVAL[25,95] **b**, Standard VGP trio assembly pipeline when DNA is available for a child and parents[20]. Dashed line indicates that the other haplotype went through the same steps before curation. In addition to the curated assemblies of both haplotypes, a representative haplotype with both sex chromosomes is submitted. **c**, Mitochondrial assembly pipeline. Figure key applies to **a**–**c**. Steps newly introduced in v1.5–v1.6 are highlighted in light blue. c, contigs; p, purged false duplications from primary contigs; q, purged alternate contigs; s, scaffolds; t, polished scaffolds. Further details and instructions are available elsewhere[33] and at https://github.com/VGP/vgp-assembly.



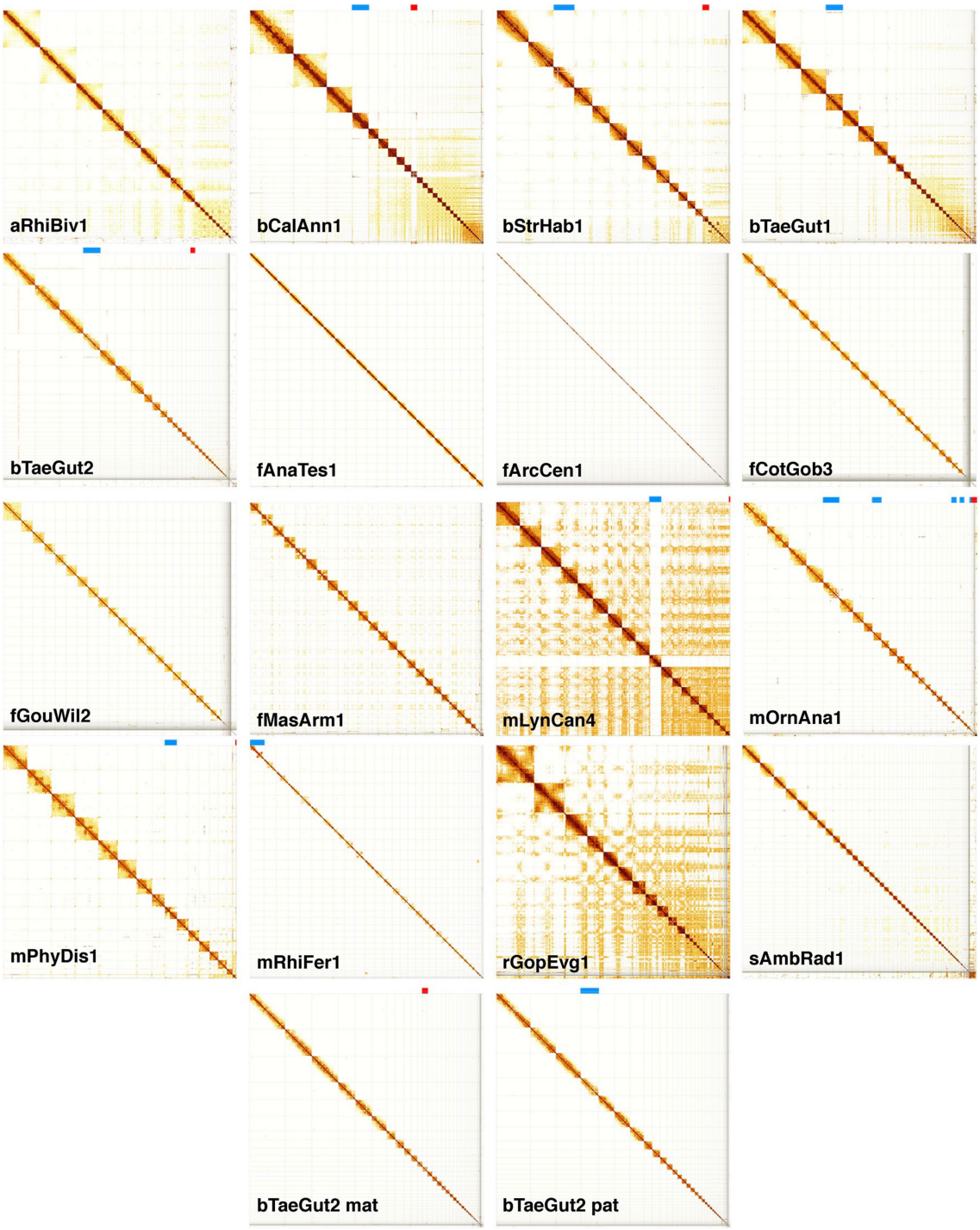
Extended Data Fig. 4 | Relationship between collapses and genomic characteristics. a, Correlation between the total number of collapses and percentage repeat content estimated in the submitted curated versions of $n = 17$ genomes from 16 species. **b**, Correlation between total number of bases in collapsed regions per Gb and repeat content. **c**, Correlation between total missing bases collapsed per Gb and repeat content. **d**, Correlation between total number of genes (coding and non-coding) in the collapsed regions and repeat content. **e**, Lack of correlation between the average collapsed size and repeat content. **f**, Lack of correlation between the total number of collapses and percentage heterozygosity. **g**, Lack of correlation between the total number of collapses per Gb and genome size. Genome size, heterozygosity, and repeat content were estimated from 31-mer counts using GenomeScope[71]. Reported are adjusted r^2 and P values from F-statistics. **h**, Cumulative collapsed bases per Gb in each collapse and percentage repeat masked. Each circle is coloured by species with its size relative to the length of the collapse as it appears in the assembly. Collapses above the horizontal bar (>90%) are further classified as collapsed high-copy repeats, and those below the horizontal bar are classified as segmental duplications (low-copy repeats). **i**, Major repeat types in collapsed high-copy repeats. Most of the repeats were masked only with WindowMasker[75], with no annotation available by RepeatMasker[104]. **j**, Minor repeat types in collapsed repeats. This is a breakdown of the repeats categorized as 'Others' in **i**, owing to the smaller scale. Bar colours in **i** and **j** are as in **h**. Note smaller scale in **j** compared with **i**. Collapsed satellite arrays were almost exclusively found in the platypus, comprising ~2.5 Mb. Collapsed simple repeats were the major source in the thorny skate (~400 kb). There was a higher proportion of LTRs in birds, LINEs and SINEs in mammals, and DNA repeats in the amphibian. Among the genes in the collapses, many were repetitive short non-coding RNAs. P values from F-statistics.



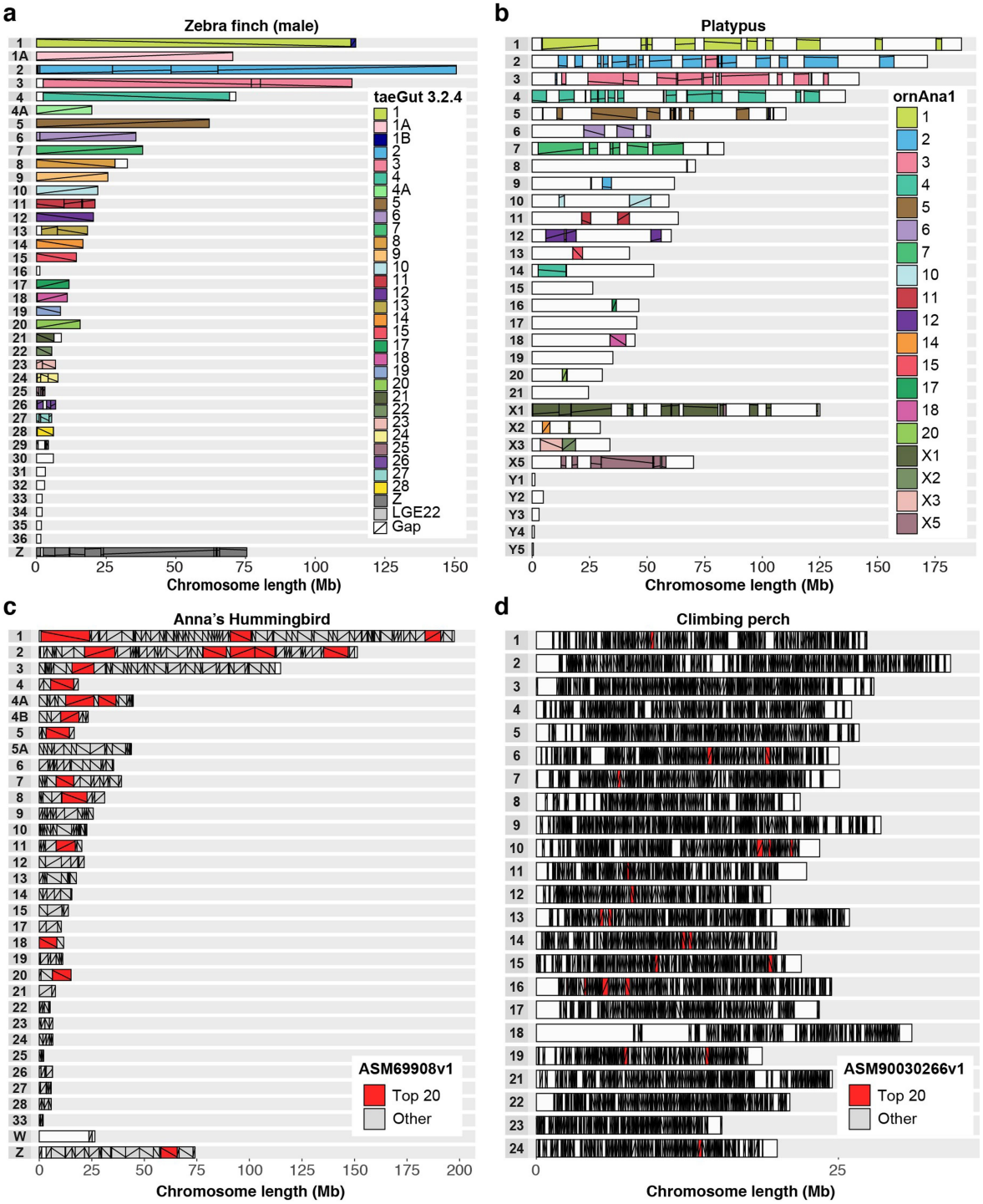
Extended Data Fig. 5 | False duplication mechanisms in genome assembly. **a**, False heterotype (haplotype) duplications occurs when more divergent sequence reads from each haplotype A (blue) and B (red) (maternal and paternal) form greater divergent paths in the assembly graph (bubbles), while nearly identical homozygous sequences (black) become collapsed. When the assembly graph is properly formed and correctly resolved (green arrow), one of the haplotype-specific paths (red or blue) is chosen for building a 'primary' pseudo-haplotype assembly and the other is set apart as an 'alternate' assembly. When the graph is not correctly resolved (purple arrow), one of four types of pattern are formed in the contigs and subsequent scaffolds. Depending on the supporting evidence, the scaffolder either keeps these haplotype contigs on separate scaffolds or brings them together on the same scaffold, often separated by gaps: 1. Separate contigs: both contigs are retained in the primary contig set, an error often observed when haplotype-specific sequences are highly diverged. 2. Flanking contigs: the assembly graph is partially formed, connecting the homozygous sequence of the 5′ side to one haplotype (blue) and the 3′ side to the other haplotype (red). 3. Partial flanking contigs: only one haplotype (blue) flanks one side of the homozygous sequence. 4. Failed connecting of contigs: all haplotype sequences fail to properly connect to flanking homozygous sequences. **b**, False homotype duplications occur where a sequence from the same genomic locus is duplicated, and are of two types: 1. Overlapping sequences at contig boundaries: in current overlap-layout-consensus assemblers, branching sequences in assembly graphs that are not selected as the primary path have a small overlapping sequence (purple), dovetailing to the primary path where it originated a branch. The size of the duplicated sequence is often the length of a corrected read. Subsequent scaffolding results in tandem duplicated sequences with a gap between. 2. Under-collapsed sequences: sequencing errors in reads (red x) randomly or systematically pile up, forming under-collapsed sequences. Subsequent duplication errors in the scaffolding are similar to the heterotype duplications. Purge_haplotigs[13] align sequences to themselves to find a smaller sequence that aligns fully to a larger contig or scaffold, and removes heterotype duplication types 1, 3, and 4. Purge_dups[14] additionally uses coverage information to detect heterotype duplication type 2 and homotype duplications. We distinguished the two types of duplications by: 1) haplotype-specific variants in reads aligning at half coverage to each heterotype duplication; 2) differing consensus quality that resulted from read coverage fluctuations when aligning reads to homotype duplications; and 3) *k*-mer copy number anomalies in which homotype duplications were observed in the assembly with more than the expected number of copies.



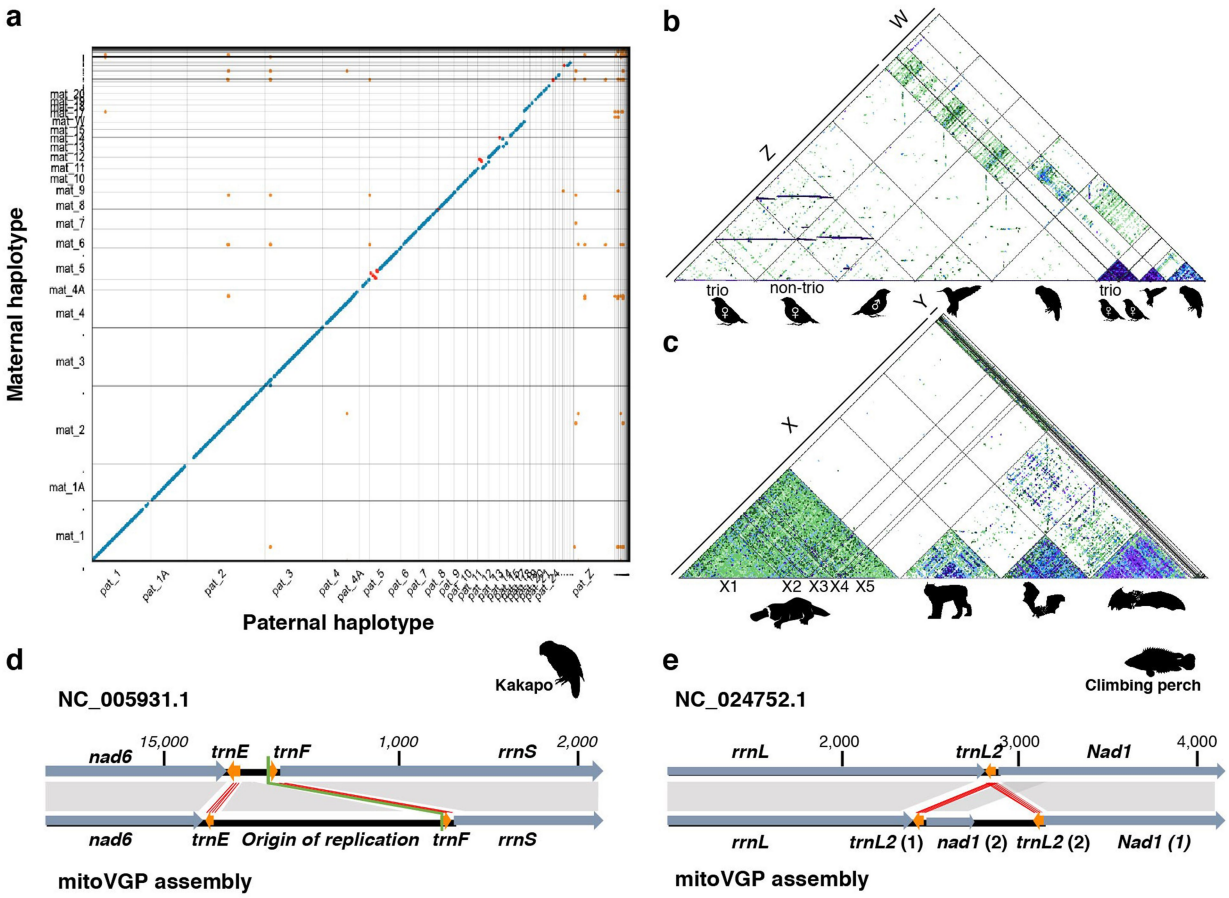
Extended Data Fig. 6 | False duplication examples fixed during manual curation. a, An example of a heterotype duplication in the female zebra finch, non-trio assembly. Left, a self-dot plot of this region generated with Gepard[105], with sequences coloured by haplotypes. Gaps, duplicated sequences (green and purple), and haplotype-specific marker densities are indicated at the top. Right, a detailed alignment view of the green haplotype duplication with paternal and maternal markers, self-alignment components, transcripts annotated, contigs, bionano maps, and repeat components displayed in gEVAL[95]. **b**, Example of a homotype duplication found in the hummingbird assembly. These were caused by an algorithm bug in FALCON, which was later fixed. **c**, Example of a combined duplication involving both heterotype (green) and homotype (orange) duplications. Assembly graph structure is shown on the left for clarity, highlighting the overlapping sites at the contig boundary shaded following the duplication type. Assembly errors including the above false duplications were detected and fixed during the curation process.



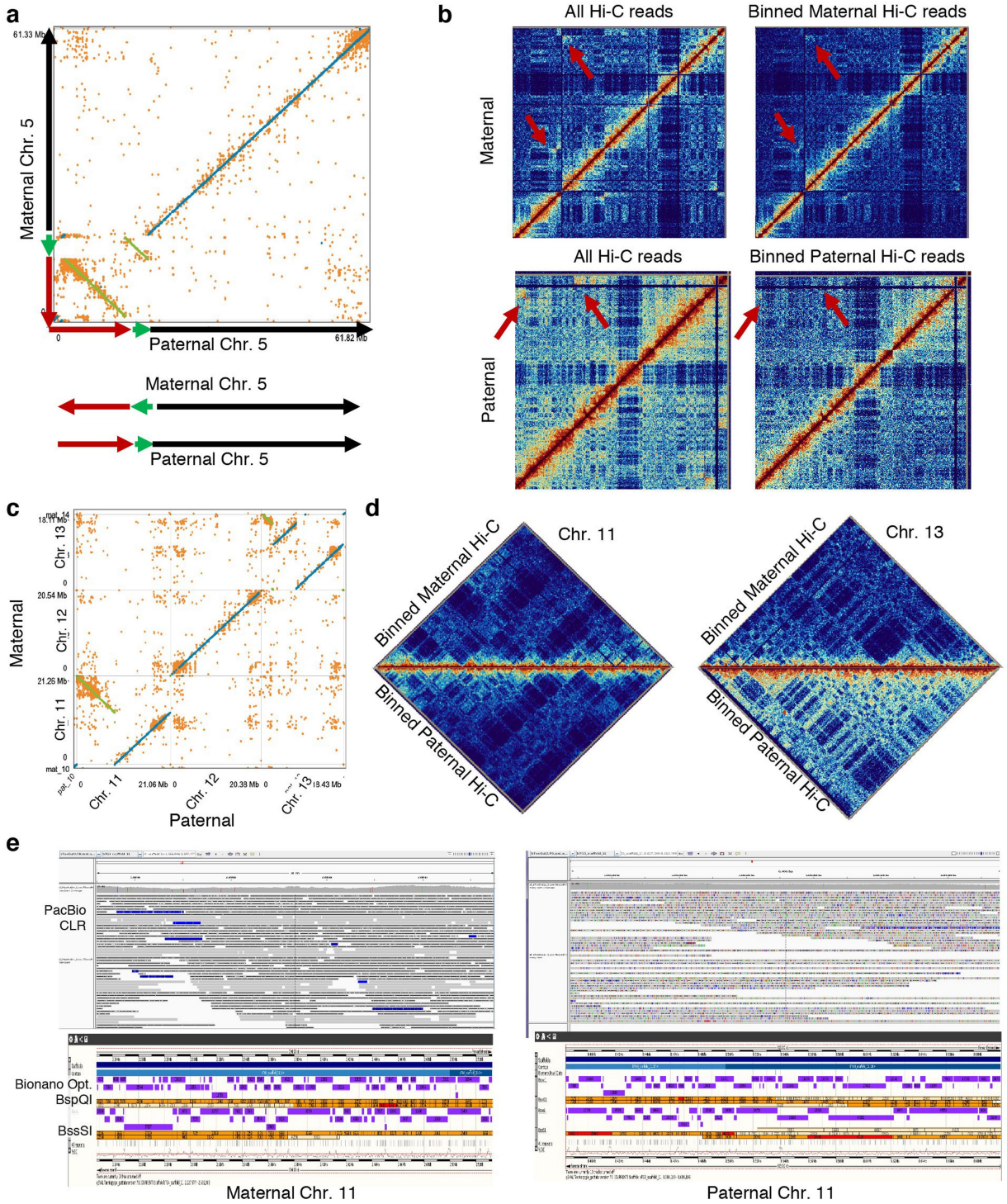
Extended Data Fig. 7 | Evidence of near-complete chromosome scaffolds in the VGP assemblies. Shown are Hi-C interaction heat maps for each species after curation, visualized with PretextView[106]. A scaffold is considered a putative arm-to-arm chromosome when all Hi-C read pairs in a row and column map to a square (that is, an assembled chromosome) on the diagonal without any other interactions off the diagonal. Those with remaining off-diagonal matches to smaller scaffolds are not linked because of ambiguous order or orientation, and are instead submitted as 'unlocalized' belonging to the relevant chromosome. Bands at the top of each heat map show scaffolds identified as X, Z (blue) or Y, W (red) sex chromosomes. The Hi-C map of fAstCal1 is not included as we had no remaining tissue left of the animal used to generate Hi-C reads.



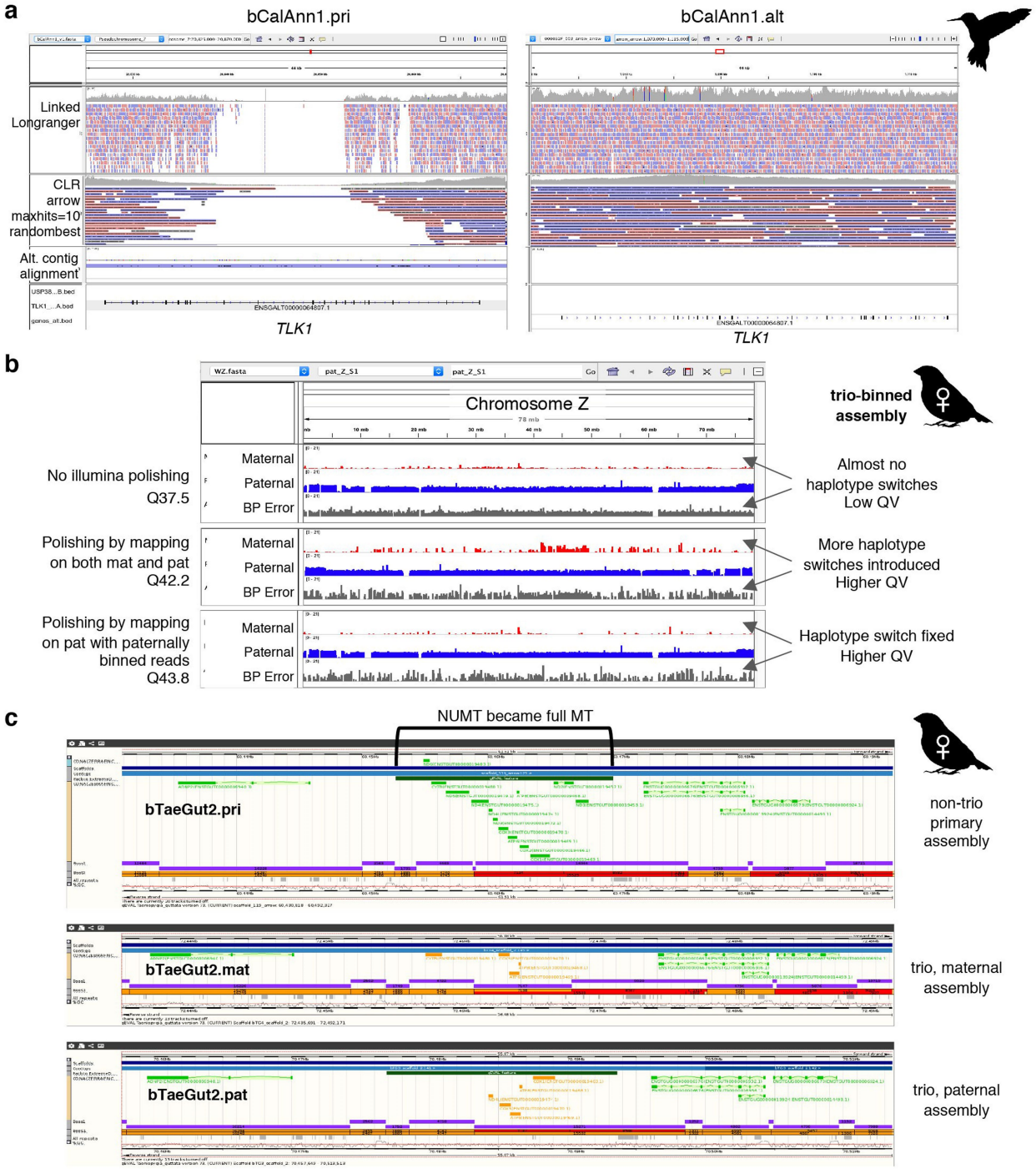
Extended Data Fig. 8 | Comparison of chromosomal organization between previous and new VGP assemblies. a, Zebra finch male compared to a previous reference assembly of the same animal. **b**, Platypus male compared with a previous reference female assembly (so the Y chromosomes are absent in the previous reference). **c**, Hummingbird female compared to a previous reference of the same animal. **d**, Climbing perch compared to a previous reference. Each row represents a VGP-generated chromosome for the target species. Colours depict identity with the reference (see key to the right); more than one colour indicates reorganization in the VGP assembly relative to the reference. The lines within each block depict orientation relative to the reference; a positive slope is the same orientation as the reference, whereas a negative slope is the inverse orientation. Gaps are white boxes with no lines, in the reference relative to the VGP assembly. A white box for the entire chromosome means a newly identified chromosome in the VGP assembly. Top 20 is the longest 20 scaffolds of the hummingbird and climbing perch assemblies. Accession numbers of the assemblies compared are listed in Supplementary Table 19.



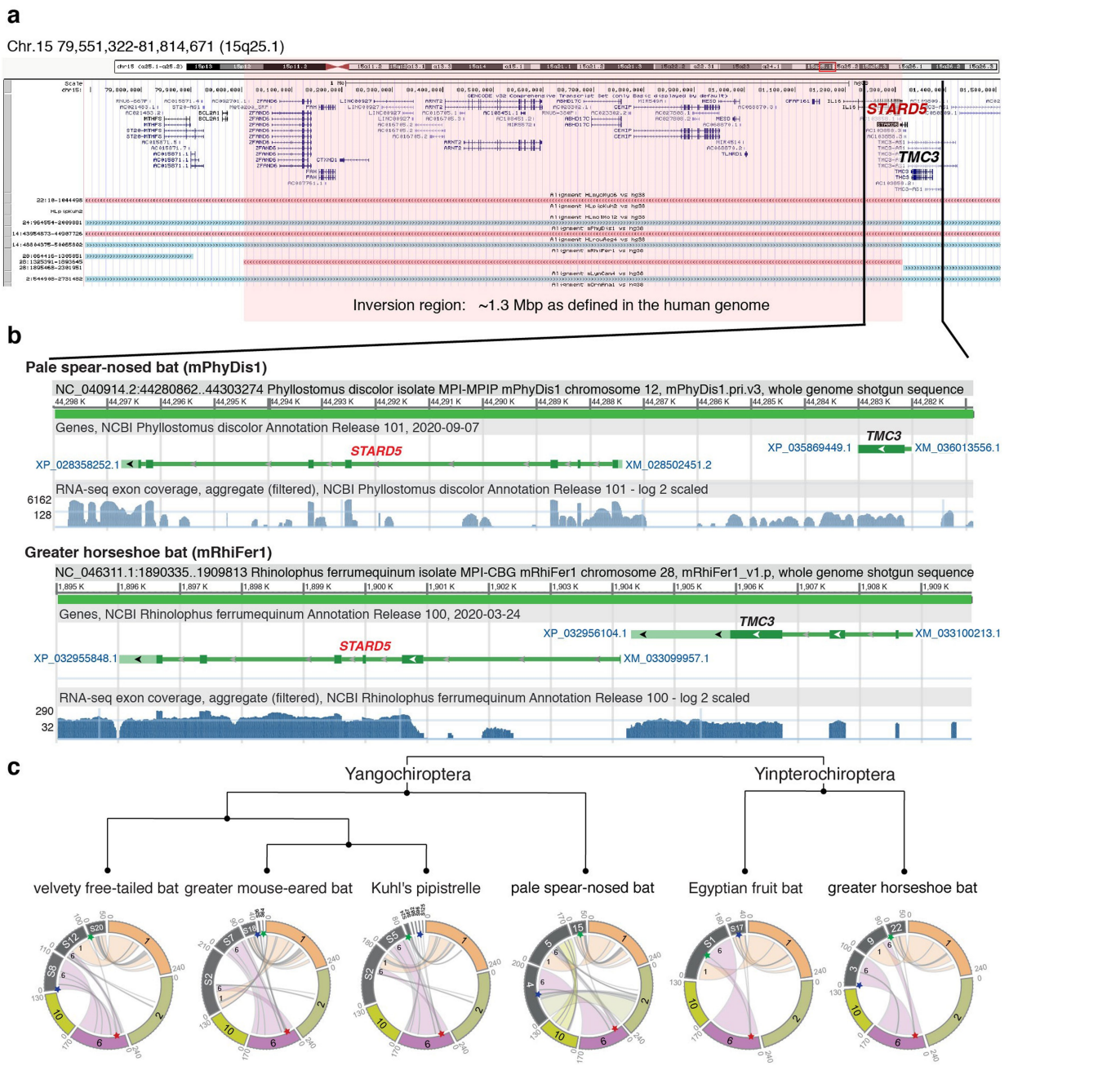
Extended Data Fig. 9 | Haplotype-resolved sex chromosomes and mitochondrial genomes. a, Alignment scatterplot, generated with MUMmer NUCmer[107], visualized with dot[108], of maternal and paternal chromosomes from the female zebra finch trio-based assembly. Blue, same orientation; red, inversion; orange, repeats between haplotypes. The paternal Z chromosome is highly divergent from the maternal W, and thus mostly unaligned. **b**, Alignment scatterplot of assembled Z and W chromosomes across the three bird species, approximated with MashMap2[86]. Segments of 300 kb (green), 500 kb (blue), and 1 Mb (purple) are shaded darker with higher sequence identity, with a minimum of 85%. The smaller size and higher repeat content of the W chromosome are clearly visible. **c**, X and Y chromosome segments of the mammals (platypus, Canada lynx, pale spear-nosed bat, and greater horseshoe bat) showing a higher density of repeats within the mammalian X chromosome than the avian Z chromosome. **d**, VGP kākāpō mitochondrial genome assembly reveals previously missing repetitive sequences (adding 2,232 bp) in the origin of replication region, containing an 83-bp repeat unit. **e**, VGP climbing perch mitochondrial genome assembly showing a duplication of *trnL2* and partial duplication of *Nad1*, which were absent from the prior reference. Orange arrows and red lines, tRNA genes and their alignments; dark grey arrows and grey shading, all other genes and their alignments; black, non-coding regions; green line, conventional starting point of the circular sequence.



Extended Data Fig. 10 | Large haplotype inversions with direct evidence in the zebra finch trio assembly. a, Two inversions (green and red) in chromosome 5 found from the MUMmer NUCmer[107] alignment of the maternal and paternal haplotype assemblies, visualized with dot[108]. **b**, Hi-C interaction plot showing that the trio-binned Hi-C data remove most of the interactions from the other haplotype (red arrows), which could be erroneously classified as a mis-assembly if only one haplotype was used as a reference. **c**, An 8.5-Mb inversion found on chromosome 11 and a complicated 8.1-Mb rearrangement on chromosome 13 between maternal and paternal haplotypes. **d**, No mis-assembly signals were detected from the binned Hi-C interaction plots, indicating that the haplotype-specific inversions are real. **e**, Half the PacBio CLR span and Bionano optical maps agree with the inversion breakpoints in chromosome 11, supporting the haplotype-specific inversion.



Extended Data Fig. 11 | Polishing artefacts. a, An example of uneven mapping coverage in the primary and alternate sequence pair of the Anna's hummingbird assembly. In this example, the alternate (alt) sequence was built at higher quality, attracting all linked-reads for polishing. The matching locus in the primary (pri) assembly was left unpolished, resulting in frameshift errors in the *TLK1* gene. **b**, Haplotype-specific markers (red for maternal, blue for paternal) and error markers found in the assembly on the Z chromosome (inherited from the paternal side) of the trio-binned female zebra finch assembly. Each row shows markers before short-read polishing, mapping all reads to both haplotype assemblies, and polishing by mapping paternally binned reads to the paternal assembly. Polishing improves QV, but introduces haplotype switch errors when using reads from both haplotypes as shown in row 2. This can be avoided when using haplotype binned reads for polishing. **c**, Example of over-polishing. The nuclear mitochondria (NuMT) sequence was transformed as a full mitochondria (MT) sequence during long-read polishing owing to the absence of the MT contig, where the NuMT attracted all long reads from the MT. In comparison, the trio-binned assembly had the MT sequence assembled in place, preventing mis-placing of MT reads during read mapping.



Extended Data Fig. 12 | Chromosome evolution among the bat species sequenced. a, Genes surrounding an inversion in the greater horseshoe bat, relative to human chromosome 15 (red highlight). The *STARD5* gene is directly disrupted by this inversion, which separates exons 1–5 from exon 6 in the greater horseshoe bat. **b**, RNA-seq tracks showing the lack of RNA splicing evidence of *STARD5* transcripts in the greater horseshoe bat (bottom) in comparison to the pale spear-nosed bat where the *STARD5* gene is not disrupted (top). **c**, Circos plots of chromosome organization relationships between the each of the analysed bats and segments of the human chromosomes 1, 2, 6 and 10. Red star, breakpoint location in human chromosome 6, depicting the fission of the boreoeutherian chromosome 5 in the bat ancestor; blue star, the region upstream of the breakpoint in the bats; green star, the region downstream of the breakpoint in the bats. The red starred breakpoint was confirmed as reused, as opposed to assembly errors, in chromosomal rearrangements of the pale spear-nosed bat, Egyptian fruit bat, and greater horseshoe bat. There is no evidence of reuse for the velvety free-tailed bat. We could not confirm breakpoint reuse in the greater mouse-eared bat or Kuhl's pipistrelle at the chromosomal scale because they were on small scaffolds that may not be completely assembled.

Extended Data Table 1 | Summary metrics of the curated and submitted vertebrate species assemblies

				Genome		Continuity					Struc. Acc.		Base Acc.			Func. Comp.	Chr. Status	
Class	Order	Species	Common name	Het (%)	Rep (%)	Prim. Size (Gb)	Alt. Size (% Prim)	Contig NG50	Scaffold NG50	Gaps / Gb	Reliable block NG50	Collapsed Mb / Gb	Map. QV	*k-mer* QV	*k-mer* Comp. (%)	BUSCO Comp. (%)	Assigned to Chr. (%)	Sex Chr.
Mammals	Chiroptera	*Phyllostomus discolor*	pale spear-nosed bat	0.9	20.3	2.1	97.3	6.4	171.7	419	15.0	3.5	39.1	38.6	96.8	93.7	99.88	XY
		Rhinolophus ferrumequinum	greater horseshoe bat	0.3	20.0	2.1	73.0	25.2	84.0	76	40.2	4.4	46.9	41.3	98.0	96.4	99.28	X
	Carnivora	*Lynx canadensis*	Canada lynx	0.2	14.7	2.4	58.8	7.4	147.3	350	23.5	1.4	39.5	36.8	97.6	94.6	99.95	XY
	Monotremata	*Ornithorhynchus anatinus*	platypus	0.4	34.6	1.9	84.7	12.4	70.1	277	17.4	7.9	42.3	40.4	96.6	96.1	98.23	X1-5 Y1-5
Birds	Passeriformes	*Taeniopygia guttata*	zebra finch (male)	1.1	11.5	1.1	91.3	12.1	71.6	295	21.5	5.7	43.4	41.5	97.2	98.2	99.25	ZW
			zebra finch (female)	1.6	13.8	1.1	85.7	4.4	73.5	713	7.2	5.4	40.8	39.4	95.4	97.8	96.08	ZW
			zebra finch (female - mat)	1.6	13.8	1.0	NA	5.4	71.7	585	12.0	NA	38.0	43.4	97.1	94.6	97.55	W
			zebra finch (female - pat)	1.6	13.8	1.0	NA	4.4	71.3	796	11.7	NA	39.1	43.7	97.1	97.9	97.16	Z
	Psittaciformes	*Strigops habroptilus*	Kākāpō	0.3	15.0	1.1	24.9	9.1	83.2	325	28.4	2.3	46.3	43.2	98.1	98.4	99.31	ZW
	Apodiformes	*Calypte anna*	Anna's hummingbird	0.6	14.4	1.1	89.9	12.8	44.7	405	33.1	3.2	40.4	38.6	94.3	96.2	99.09	ZW
Reptiles	Testudines	*Gopherus evgoodei*	Goode's thornscrub tortoise	0.4	30.6	2.3	85.3	10.5	131.6	245	17.1	8.7	41.1	38.4	96.2	96.2	94.92	-
Amphibians	Rhinatrematidae	*Rhinatrema bivittatum*	two-lined caecilian	0.5	33.0	5.3	86.8	3.4	486.9	672	6.1	0.7	38.7	38.2	96.7	81.9	97.36	-
Teloest Fishes	Synbranchiforme	*Mastacembelus armatus*	zig-zag eel	0.1	33.2	0.6	7.4	4.8	23.3	402	10.6	9.6	38.7	44.5	95.9	98.0	96.77	-
	Anabantiformes	*Anabas testudineus*	climbing perch	0.6	24.6	0.6	97.0	4.6	23.5	479	13.0	2.5	24.3	37.1	92.6	98.4	99.75	-
	Cichliformes	*Archocentrus centrarchus*	flier cichlid	0.4	29.4	0.9	110.8	2.0	35.6	797	2.6	6.2	38.4	34.7	94.5	97.3	88.79	-
		Astatotilapia calliptera	eastern happy cichlid	0.2	30.8	0.9	No alt	4.0	36.7	557	2.8	11.6	34.8	37.4	93.9	97.1	96.62	-
	Perciformes	*Cottoperca trigloides (gobio)*	channel bull blenny	0.3	28.7	0.6	110.2	5.9	25.2	730	3.4	5.5	28.8	33.6	87.2	91.8	94.36	-
	Gobiesociformes	*Gouania adriatica (willdenowi)*	blunt-snouted clingfish	0.5	43.1	0.9	6.7	1.2	36.8	1,238	3.9	17.5	33.9	38.6	87.9	95.4	93.54	-
Cartilaginous fishes	Rajformes	*Amblyraja radiata*	thorny skate	0.9	54.1	2.6	57.2	0.8	49.8	1,467	2.3	31.4	38.4	39.3	87.8	88.6	95.37	X

Colour shading indicates degree of heterozygosity or repeats (red), primary assembly sizes and relative size of alternate haplotypes (orange), continuity measures (green), gaps and collapses (blue), and base call accuracy (purple). A dash indicates that the sex chromosomes were not found or are not known. Accessions are available in Supplementary Table 10.

Extended Data Table 2 | Annotation summary statistics in previous and newly assembled VGP reference genomes

Species		Hummingbird		Zebra finch			Platypus	
	Assembly	ASM69908v1	bCalAnn1_v1.p (VGP)	Taeniopygia_guttata-3.2.4	bTaeGut1_v1.p (VGP)	bTaeGut2.pat.v3 +W (VGP)	Ornithorhynchus_anatinus-5.0.1	mOrnAna1.p.v1 (VGP)
	Assembly accession	GCF_000699085.1	GCF_003957555.1	GCF_000151805.1	GCF_003957565.1	GCF_008822105.2	GCF_000002275.2	GCF_004115215.1
	NCBI Annotation Release	N/A	101	N/A	104	105	N/A	104
All Genes	Total	16,234	16,230	24,719	22,186	21,543	34,289	30,932
	Genes with alternative variants	3,649	5,897	6,429	9,312	9,130	4,414	9,485
Coding genes	Total	14,714	14,711	18,482	17,438	16,197	19,883	18,200
	With orthologs to human	12,250	12,502	11,709	12,801	12,903	10,563	14,730
	> 80% cov. by SwissProt protein	13,330	13,538	15,569	15,154	14,090	16,165	16,486
	> 80% SwissProt cov. by protein	10,697	13,237	11,661	15,845	14,830	8,635	16,137
Problematic coding genes	**Partial**	**1,637**	**110**	**4,109**	**212**	**289**	**5,601**	**228**
	with >5% *ab initio*	1,828	890	1,557	702	534	4,676	1,334
	with corrections	909	1,056	2,658	739	468	1,258	834
CDS	**Total**	**22,448**	**29,231**	**36,951**	**43,977**	**42,385**	**29,130**	**43,203**
	Fully-supported (no *ab initio*)	18,280	27,549	33,879	42,799	41,466	20,630	40,677
	Mean length	1,774	1,949	1,885	2,263	2,283	1,494	2,077
	Median length	1,332	1,452	1,334	1,620	1,623	1,047	1,512

Annotation results of VGP assemblies and previous reference assemblies with the NCBI Eukaryotic Genome Annotation Pipeline, using the same RNA-seq data and nearly identical sets of transcripts and proteins on input. Highlighted rows are plotted in Fig. 5c, d.

nature research

Corresponding author(s): Erich D. Jarvis, Arang Rhie, Kerstin Howe, Eugene Myers, Richard Durbin, Adam Phillippy

Last updated by author(s): Jan 8, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We collected genome sequence data of various technologies from individuals representing 16 vertebrate species.
Data analysis	All software used is at at https://github.com/VGP/vgp-assembly, as of November 2020. The assembly software and versions used are listed in Supplementary Table The specific VGP assembly pipeline used is at https://github.com/VGP/vgp-assembly/tree/master/pipeline.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw sequence and assembly data are available at Genome Ark https://vgp.github.io/genomeark/ and the NCBI BioProject page PRJNA489243 https://www.ncbi.nlm.nih.gov/bioproject/489243

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For most analyses, we analyzed data from a sample size of n = 16 vertebrate species, and n = 17 individuals; we had two individuals for one species (the zebra finch). For the chromosomal evolution analyses, we added an additional n = 4 bat species.
Data exclusions	Only sequence data that failed quality control were excluded or repeated.
Replication	We confirmed the ability to replicate all code using multiple rounds of assembly or analyses.
Randomization	No analyses required generating randomize data sets.
Blinding	No analyses required being blind to groups

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	The care and collection of a laboratory female zebra finch sample was done under an approved IACUC protocol at the Rockefeller University. The male laboratory sample was obtained from the approval mention of the previous reference genome in Warren et al 2011 Nature.
Wild animals	Samples of the 15 other species were collected from wild animals, with approved permits of the source institutions and local governments involved. These sources, persons with the permits, geographic location, sex and relative age are listed in Supplementary Table 8 and the BioSample submissions in NCBI and ENA.
Field-collected samples	No laboratory work was conducted on field-collected animals
Ethics oversight	Rockefeller University for the zebra finch species.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Chapter 2

Secomandi et al. **"Pangenomics provides insights into the role of synanthropy in barn swallow evolution"**. *Under review, Cell Genomics*.[9]

[9]Supplementary materials can be found on the preprint bioRxiv page: Pangenomics provides insights into the role of synanthropy in barn swallow evolution | bioRxiv

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.

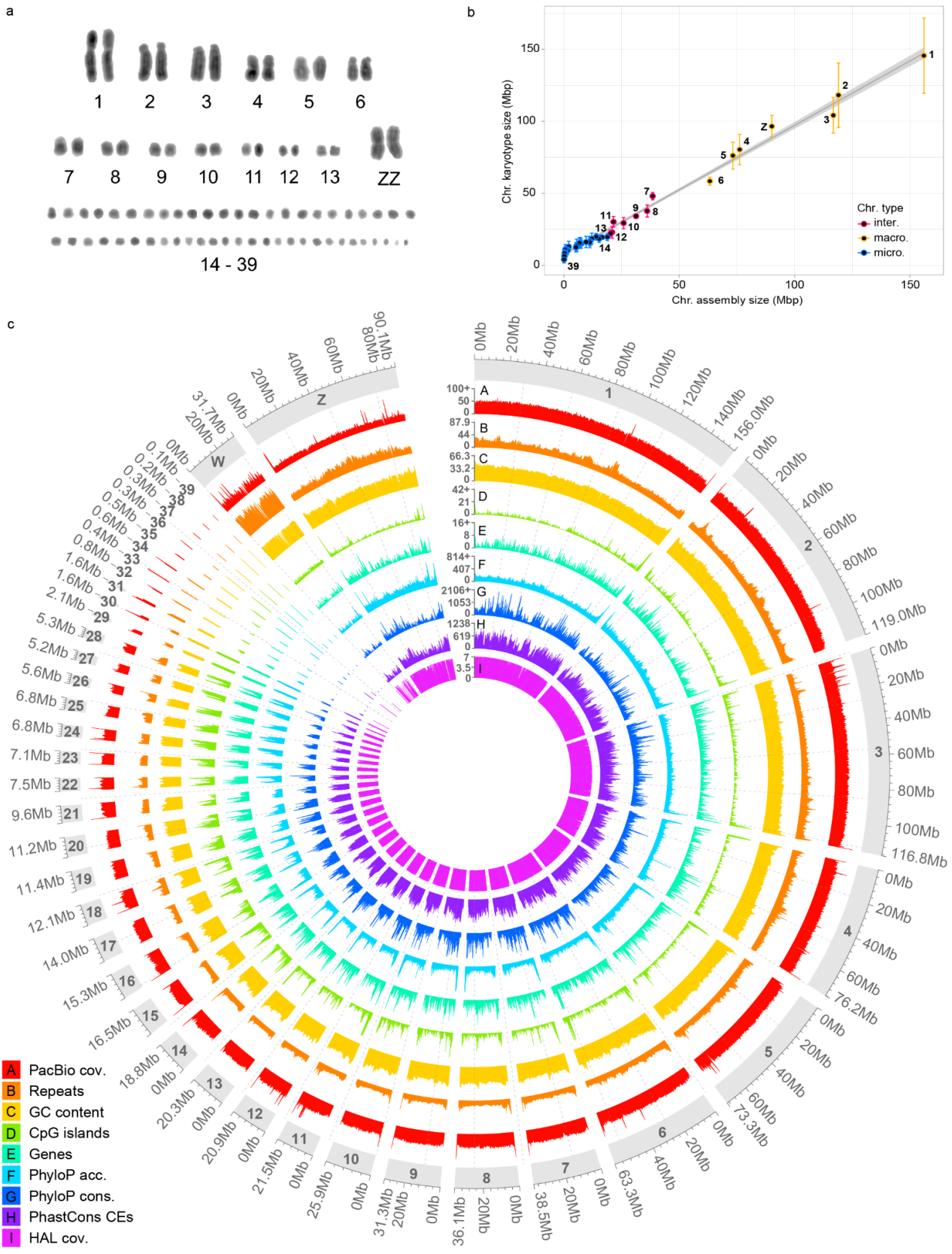


Fig. 1 Karyotype reconstruction and chromosome-level assembly for the barn swallow. **a** DAPI-stained karyotype of a male Eurasian barn swallow (inverted colours). **b** Correlation between assembled chromosome length (x) and the estimated chromosome length from karyotype images (y). W sex chromosome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

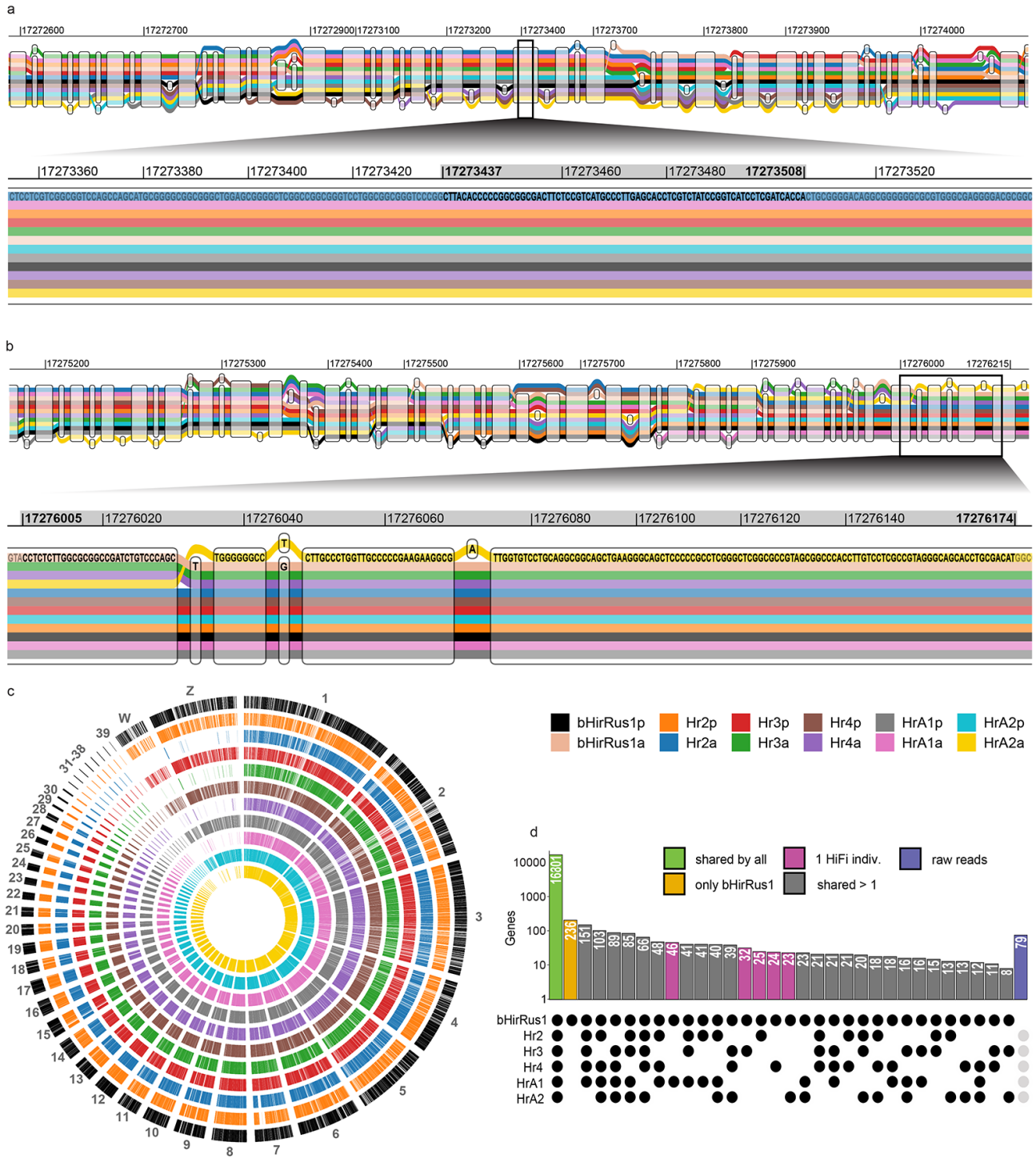


Fig. 2 The first pangenome for the barn swallow. **a** *CAMK2N2* initial region in the barn swallow pangenome. bHirRus1 Chr10 ('bHirRus1p') is shown together with the alternate assembly 'bHirRus1a', the five HiFi-based primary (Hr2p, Hr3p, Hr4p, HrA1p, HrA2p), and their alternate assemblies (Hr2a, Hr3a, 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).

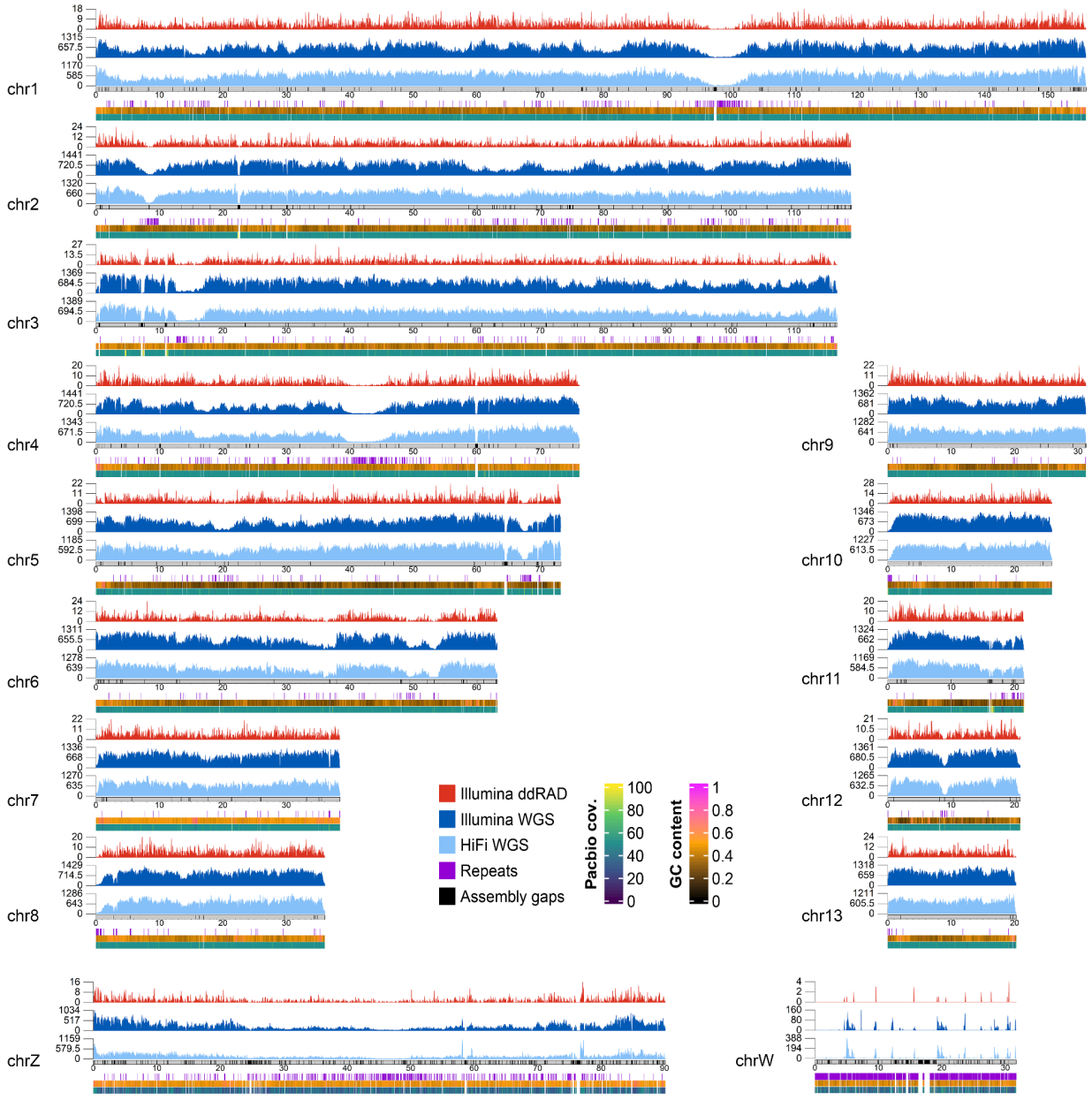


Fig. 3 SNPs density per chromosome. Macrochromosomes and intermediate chromosomes are shown here, while microchromosomes are shown in Extended Data Fig. 7. SNP density, coloured according to the different types of genomic data used, was computed over 40 kbp windows. The numbers on the y axis of each 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).

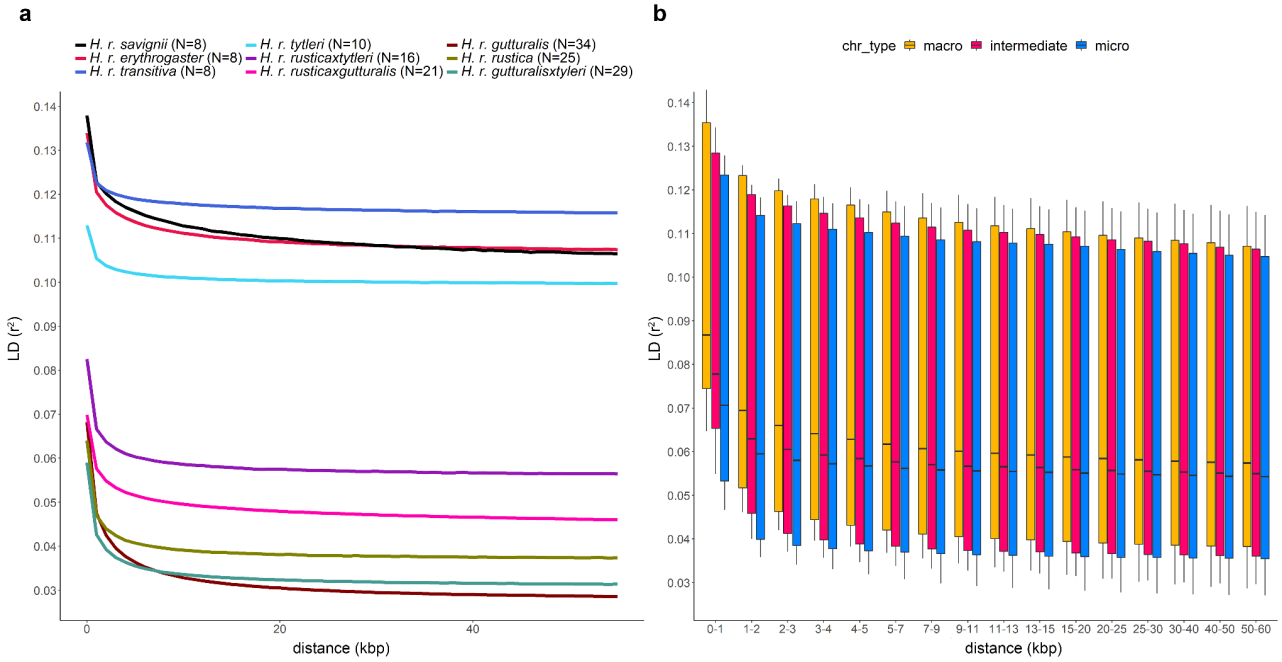


Fig. 4 Linkage disequilibrium decay in the barn swallow genome. **a** Average r^2 values plotted against physical distance (kbp) for the different populations belonging to ds2 and ds3.1 (Illumina WGS data). **b** Average r^2 values in macrochromosomes, intermediate chromosomes, and microchromosomes according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

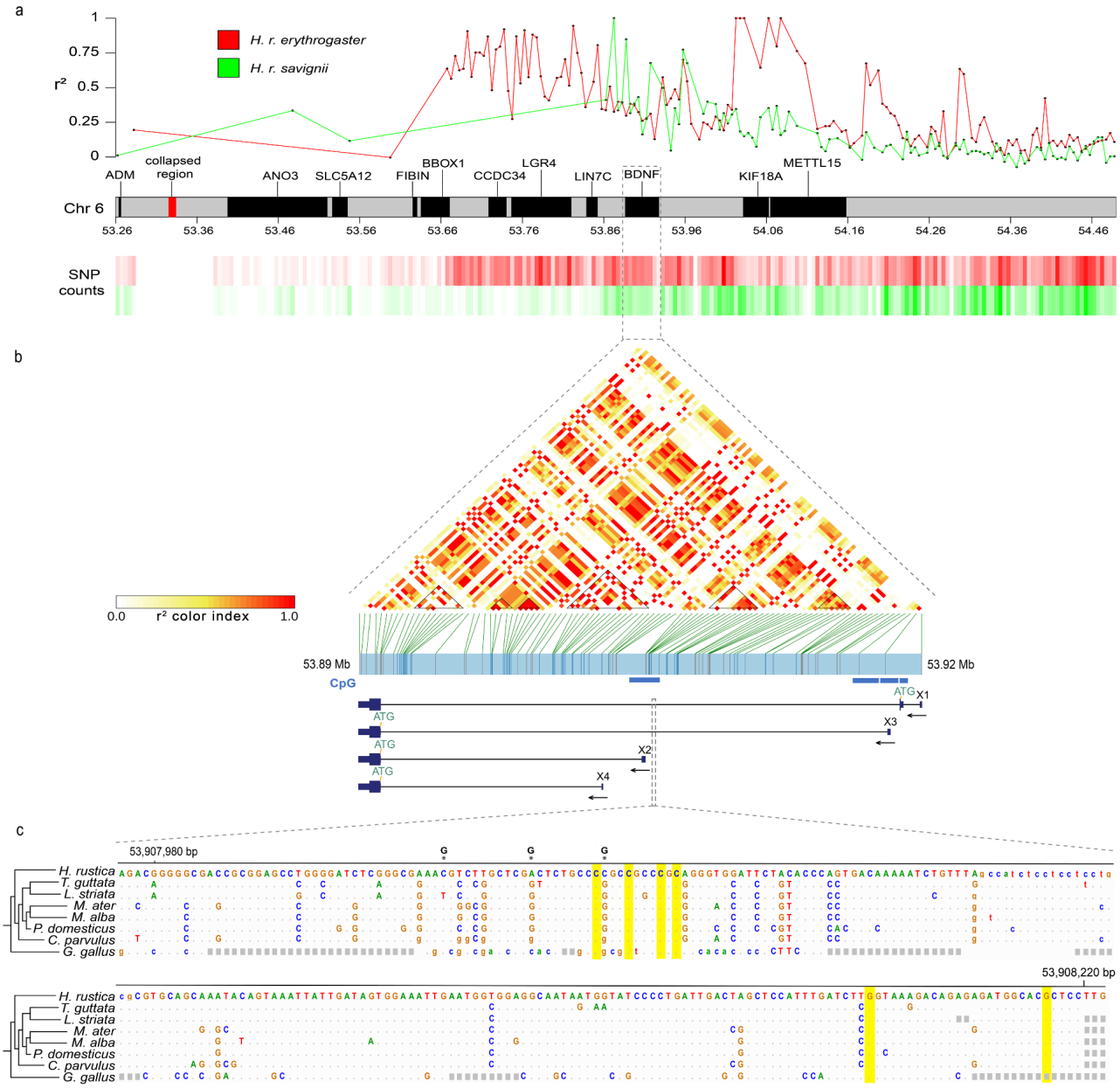


Fig. 5 Patterns of LD blocks in genomic regions on chr6. **a** Average r^2 values computed over 5 kbp windows on chr. 6 (upper panel; from 53.26 Mb to 54.49 Mb) for the *H. r. savignii* (green) and *H. r. erythrogaster* (red) populations (ds3.1). The region shown in the plot extends beyond ROI 45. Each 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.

Chapter 3

Lombardo et al. (2022) "**The mitogenome relationships and phylogeography of barn swallows (*Hirundo rustica*)**". *Accepted manuscript, Molecular Biology and Evolution*.

Article (Discoveries)

The Mitogenome Relationships and Phylogeography of Barn Swallows (*Hirundo rustica*)

Gianluca Lombardo,[1] Nicola Rambaldi Migliore,[1] Giulia Colombo,[1] Marco Rosario Capodiferro,[1] Giulio Formenti,[2] Manuela Caprioli,[3] Elisabetta Moroni,[1] Leonardo Caporali,[4] Hovirag Lancioni,[5] Simona Secomandi,[6] Guido Roberto Gallo,[6] Alessandra Costanzo,[3] Andrea Romano,[3] Maria Garofalo,[7] Cristina Cereda,[7] Valerio Carelli,[4,8] Lauren Gillespie,[9] Yang Liu,[10] Yosef Kiat,[11] Alfonso Marzal,[12] Cosme López-Calderón,[13] Javier Balbontín,[14] Timothy A. Mousseau,[15] Piotr Matyjasiak,[16] Anders Pape Møller,[17] Ornella Semino,[1] Roberto Ambrosini,[3] Andrea Bonisoli Alquati,[18] Diego Rubolini,[3] Luca Ferretti,[1] Alessandro Achilli,[1] Luca Gianfranceschi,[6] Anna Olivieri,[*,1] and Antonio Torroni[*,1]

[1] Dipartimento di Biologia e Biotecnologie "Lazzaro Spallanzani", Università di Pavia, 27100 Pavia, Italy

[2] Vertebrate Genome Laboratory, The Rockefeller University, New York, NY 10065, USA

[3] Dipartimento di Scienze e Politiche Ambientali, Università degli Studi di Milano, 20133 Milan, Italy

[4] IRCCS Istituto delle Scienze Neurologiche di Bologna, Programma di Neurogenetica, 40139 Bologna, Italy

[5] Dipartimento di Chimica, Biologia e Biotecnologie, Università di Perugia, 06123 Perugia, Italy

[6] Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milan, Italy

[7] Genomic and Post-Genomic Unit, IRCCS Mondino Foundation, 27100 Pavia, Italy

[8] Dipartimento di Scienze Biomediche e Neuromotorie, Università di Bologna, 40139 Bologna, Italy

[9] Department of Academic Education, Central Community College, Columbus, NE 68601, USA

[10] State Key Laboratory of Biocontrol, School of Ecology, Sun Yat-sen University, Guangzhou 510275, China

[11] Israeli Bird Ringing Center (IBRC), Israel Ornithological Center, Tel Aviv, Israel

[12] Department of Zoology, University of Extremadura, 06071 Badajoz, Spain

[13] Department of Wetland Ecology, Estación Biológica de Doñana CSIC, 41092 Seville, Spain

[14] Department of Zoology, University of Seville, 41012 Seville, Spain

[15] Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA

[16] Institute of Biological Sciences, Cardinal Stefan Wyszyński University in Warsaw, 01-938 Warsaw, Poland

[17] Ecologie Systématique Evolution, Université Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91405, Orsay Cedex, France

[18] Department of Biological Sciences, California State Polytechnic University - Pomona, Pomona, CA 91767, USA

[*] Shared last authors and **corresponding authors:** E-mails: antonio.torroni@unipv.it; anna.olivieri@unipv.it

Key words: barn swallow phylogeny, *Hirundo rustica* subspecies, mitogenome, haplogroups.

Abstract

The barn swallow (*Hirundo rustica*) poses a number of fascinating scientific questions, including the taxonomic status of postulated subspecies. Here we obtained and assessed the sequence variation of 411 complete mitogenomes, mainly from the European *H. r. rustica*, but other subspecies as well. In almost every case, we observed subspecies-specific haplogroups, which we employed together with estimated radiation times to postulate a model for the geographical and temporal worldwide spread of the species. The female barn swallow carrying the *Hirundo rustica* ancestral mitogenome left Africa (or its vicinity) around 280 thousand years ago (kya), and her descendants expanded first into Eurasia and then, at least 51 kya, into the Americas, from where a relatively recent (< 20 kya) back migration to Asia took place. The exception to the haplogroup subspecies specificity is represented by the sedentary Levantine *H. r. transitiva* that extensively shares haplogroup A with the migratory European *H. r. rustica* and, to a lesser extent, haplogroup B with the Egyptian *H. r. savignii*. Our data indicate that *rustica* and *transitiva* most likely derive from a sedentary Levantine population source that split at the end of the Younger Dryas (11.7 kya). Since then, however, *transitiva* received genetic inputs from and admixed with both the closely related *rustica* and the adjacent *savignii*. Demographic analyses confirm this species' strong link with climate fluctuations and human activities making it an excellent indicator for monitoring and assessing the impact of current global changes on wildlife.

Introduction

The barn swallow (*Hirundo rustica*) is one the most widely distributed bird species (Turner and Rose 2010), possibly due to the switch from natural nesting sites, especially caves, to nesting in human-made structures (Zink et al. 2006). This commensal and iconic species for numerous human groups and cultures is portrayed in art, myths, legends and poetry for millennia (Green 1988), and comprises at least six subspecies, all with breeding ranges in the Holarctic (but see Areta et al. 2021). The subspecies differ in several morphometric characteristics, such as body size, length of outer tail streamers, ventral coloration, and extent of the dark breast band (Turner 2006). The subspecies include *H. r. rustica* (Europe, North Africa and Western Asia), *H. r. savignii* (Egypt), *H. r. transitiva* (Israel, Lebanon, Jordan and Syria), *H. r. tytleri* (southern-central Siberia, Mongolia), *H. r. gutturalis* (central-eastern China, Japan) and *H. r. erythrogaster* (North America). Additional subspecies such as *H. r. saturata* and *H. r. mandshurica* have been postulated in north-eastern Asia, but their distinct subspecies status relative to the other Asian subspecies is debated (Cheng 1987; Brown and Brown, 1999; Dickinson and Dekker, 2001; Dickinson et al. 2002; Turner 2006; Liu et al. 2020). While the *Hirundo rustica* species complex is not endangered, local populations or even subspecies show declines due to specific threats, mostly related to agricultural intensification (Ambrosini et al. 2012; Møller 2019). Most subspecies are migratory, and their wintering grounds cover much of the southern hemisphere as far south as central Argentina, the Cape province of South Africa, and northern Australia (Turner 2006; Hobson et al. 2015; Liechti et al. 2015; Winkler et al. 2017). Adult swallows are highly philopatric (Møller 1994), whereas natal dispersal is relatively large, with some individuals, especially females, dispersing up to several hundreds of kilometers from their natal site (Turner 2006; Balbontín et al. 2009; Scandolara et al. 2014). However, *H. r. savignii* and *H. r. transitiva* are sedentary throughout the year (Shirihai et al. 1996; Turner 2006; Turner and Rose 2010), or make short-distance movements during the non-breeding period (Kiat, unpublished data).

The earliest study of barn swallow nuclear DNA variation (*MUSK* gene) did not detect a genetic structure within the species, suggesting a rather recent subspecies differentiation (Zink et al. 2006). More recent and extensive surveys of microsatellite and ddRAD sequence data in *H. r. rustica* revealed a lack of population structure among breeding populations from Sweden, Germany and Switzerland with no evidence of genomic selection between phenotypic migratory types (Santure et al. 2010; von Rönn et al. 2016). In contrast, genotyping of over 9,000 SNPs in 350 barn swallows from four subspecies revealed genome-wide clustering that generally corresponds with the subspecies, a certain level of differentiation of the UK population (*H. r. rustica)* from eastern European and Turkish populations of the same subspecies, and genomic covariance of the latter *H. r. rustica* populations with non-migratory *H. r. transitiva* specimens from Israel (Safran et al. 2016). With a similar approach, molecular evidence of hybridization between subspecies was also obtained (Scordato et al. 2017; 2020).

In the last few years, whole genome sequencing data have been reported for a few subspecies (*H. r. erythrogaster, H. r. savignii*) (Safran et al. 2016; Smith et al. 2018), including the first reference genome draft (*H. r. rustica*) (Formenti et al. 2019). Recently, in the framework of the Vertebrate Genomes Project, an effort to generate complete and accurate genome assemblies for all vertebrate species, a new reference genome for *H. r. rustica* as well as the first pangenome for the species was released. This allowed the assessment of the extent of conservation and acceleration in the barn swallow genome and the identification of a catalogue of genetic markers and candidate genomic regions under selection (Formenti G., data not shown).

So far, however, most genetic studies concerning the relationships between barn swallow subspecies have focused on the maternally-transmitted and fast-evolving mitochondrial DNA (mtDNA), particularly on the sequence variation of single mitochondrial genes, such as *ND2* and *CYB* (Sheldon et al. 2005; Zink et al. 2006; Dor et al. 2010, 2012; Malaitad et al. 2016). They

confirmed that the barn swallow species complex is monophyletic, and revealed that the different subspecies cluster into two major phylogenetic branches, which might have diverged approximately 100 thousand years ago (kya) and geographically correspond to Europe-Middle East and Asia-America (Zink et al. 2006; Dor et al. 2010), thus substantially predating human agriculture and the new nesting opportunities provided by human settlements. Moreover, the close relationship between one of the Asian subspecies (*H. r. tytleri*) and the American one (*H. r. erythrogaster*) has raised the possibility of a secondary dispersal event, possibly about 27 kya, from North America back into Asia (Zink et al. 2006). Finally, the potential lack of differentiation between the migratory *H. r. rustica* and the sedentary *H. r. transitiva* was also observed with the fast-evolving mtDNA (Dor et al. 2010), suggesting intermingling between the two subspecies.

Despite the valuable genetic insights provided by these studies, the assessment of only a rather short segment of the barn swallow mtDNA limits their phylogenetic resolution, and the understanding of this species' origin and spread. A finer phylogenetic resolution can be achieved by sequencing the entire mitogenome, an approach that has been employed for humans and many other animal species (Achilli et al. 2008, 2012; Behar et al. 2012; Miao et al. 2013; Morin et al. 2015; Battaglia et al. 2016; Barth et al. 2017; Peng et al. 2018; Cole et al. 2019; de Manuel et al. 2020; Niedziałkowska et al. 2021), and recently pursued also in *H. rustica* (Carter et al. 2020). Here, we exploited next generation sequencing (NGS) to obtain 411 complete mitogenomes, mainly from the European *H. r. rustica*, but also from other subspecies. Phylogenetic and Bayesian analyses allowed us to (i) obtain a high-resolution mitogenome phylogeny of the species, (ii) better define the matrilineal relationships and links between subspecies and their divergence times, and (iii) assess demography through time.

Results and Discussion

Organization of the Barn Swallow Mitogenome

Our first complete mitogenome was obtained from a *H. r. rustica* specimen (#20) from Italy (supplementary table S1, Supplementary Material online). This mitogenome (MZ905359), employed as *H. r. rustica* Reference Sequence (HrrRS), was Sanger sequenced together with four additional *H. r. rustica* mitogenomes from Italy (#1, 35, 136 and 302). The mitogenome is 18,143 bps in length and harbours 37 genes: 13 protein-coding, 22 tRNA, and two rRNA genes, as well as two non-coding regions, CR1 and CR2, following the GO-II gene order (Mackiewicz et al. 2019; Urantówka et al. 2020) (supplementary fig. S1 and supplementary table S2, Supplementary Material online).

NGS technology was employed to sequence additional 405 entire barn swallow mitogenomes and another was extrapolated from Formenti et al. (2019). These mitogenomes were from four putative subspecies (336 *H. r. rustica*, 50 *H. r. transitiva*, 5 *H. r. gutturalis* and 15 *H. r. erythrogaster*); for *H. r. rustica*, they were from numerous sampling locations (fig. 1, supplementary fig. S2 and supplementary table S1, Supplementary Material online). A total of 387 distinct haplotypes were detected, with 1,385 variable sites in the coding region (15,601 bps; nps 1-14859, nps 16068-16740, nps 18075-18143). On average, 32.8 ± 1.0 nucleotide differences were found between any two coding-region sequences. The average π value for the 411 entire mitogenomes is 0.226% (± 0.018%) with the highest variability in the two control regions (Supplementary fig. S3, Supplementary Material online). A total of 1102 synonymous and 156 non-synonymous mutations were identified in the 13 protein-coding genes (PCGs) (supplementary fig. S4, Supplementary Material online). As expected, all loci harbour more synonymous than non-synonymous mutations indicating the action of purifying selection (Stewart et al. 2008).

The Phylogeography of Barn Swallow Mitogenomes and Haplogroup Ages

Phylogenetic analyses reveal that all 411 *Hirundo rustica* mitogenomes cluster into four main branches that we named haplogroups A, B, C and D (fig. 1 and supplementary fig. S5,

Supplementary Material online). These mitogenomes derive from a common female ancestor that harboured the *H. rustica* ancestral mitogenome (HrAM). Consistent with previous results, the four haplogroups are included in two primary branches (Zink et al. 2006; Dor et al. 2010) resulting from the first split in the phylogeny. One of the branches includes haplogroups A and B and the other encompasses haplogroups C and D. We thus named them AB and CD, respectively. As previously noted (Dor et al. 2010), this division is supported by a plumage trait, the dark breast band, which is broad and complete in the subspecies clustering within the AB branch (*H. r. rustica*, *H. r. transitiva*, *H. r. savignii)*, and narrow or incomplete in those with CD mitogenomes (*H. r. gutturalis*, *H. r. erythrogaster*, *H. r. tytleri*).

For all nodes in the phylogeny and the derived haplogroups and sub-haplogroups, we obtained age estimates both with Maximum Likelihood (ML) and Bayesian approaches. The estimates obtained with the two methods are very similar (supplementary table S3, Supplementary Material online). Thus, for brevity we report here only the Bayesian ages.

According to our data, the female barn swallow carrying the HrAM lived 276.9 ± 24.3 kya, an almost three-fold age increase relative to earlier estimates (Zink et al. 2006; Dor et al. 2010). A result of this type was not unexpected. Indeed, by improving the molecular and phylogenetic resolution of mtDNA to the level of entire (or almost entire) mitogenomes, important age changes for the most recent common female ancestor were reported in different species (Achilli et al. 2012), including humans (Torroni et al. 2006; Behar et al. 2012).

Of the four main haplogroups, haplogroup A is by far the best represented (n = 388) in our sample (figs. 1 and 2). It began to radiate 57.1 ± 6.4 kya and comprises all mitogenomes from Europe and Algeria (*H. r. rustica*) as well as 46 of the 50 *H. r. transitiva* mitogenomes from Israel and one from *H. r. gutturalis* (#258) sampled in China (Nujiang Prefecture, Yunnan Province). Three sub-haplogroups originated from its initial split, the largely predominant A1 and the rare A2 and A3, with the former harbouring two very common sub-branches detected in all European locations as well as in Algeria (fig. 1), with mitogenomes from each location generally scattered and intermingled with those from the other European locations. Furthermore, we observed that the 46 mitogenomes from *H. r. transitiva* belonging to A (black dots in fig. 2) are also scattered among the *H. r. rustica* mitogenomes.

These observations tend to confirm the rather poor genetic differentiation of *H. r. rustica* populations at a high level of molecular and phylogenetic resolution, and of *H. r. transitiva* too, at least for the predominant haplogroup A component. Our *H. r. transitiva* sample from Israel would be essentially indistinguishable from the European *H. r. rustica* populations, if not for the detection of four haplogroup B (8.0%) mitogenomes (fig. 1, supplementary table S1 and supplementary fig. S5, Supplementary Material online). A diffuse and broad overlap of the mtDNA variation between *H. r. rustica* and *H. r. transitiva* is also confirmed by the haplogroup A diversity values in the two subspecies, which are identical (0.13%) (supplementary table S4, Supplementary Material online). Three possible explanations can be envisioned for the extensive mtDNA overlap between *rustica* and *transitiva*. First, the two adjacent subspecies derive from the same ancestral source in which A was the only (or predominant) haplogroup and was already differentiated into sub-haplogroups at the time of the initial *rustica-transitiva* split. Alternatively, *rustica* and *transitiva* maternal lineages underwent gene flow, possibly continuously over time. Finally, *rustica* and *transitiva* derive from the same ancestral population, but also admix; a process that is still going on, despite the (growing) differences in migratory behaviour, moult strategy (Kiat et al. 2019) and morphology, when migrant *rustica* individuals pass through the *transitiva* breeding areas at the main time of their breeding season.

Nevertheless, because of the abundance of haplogroup A mitogenomes in our collection, we also detected a certain amount of genetic differentiation among populations. Indeed, a number of

subclades harbour rather localized geographic distributions and appear to be population-specific. These subclades are not uncommon and sometimes they are relatively ancient: four were found in Spain (2-3 haplotypes each) with the oldest (A1a1b3b) dating ~11.4 ky; 20 in Italy (2-5 haplotypes each) with the oldest (A1a2g) dating ~11.6 ky; one (A1a2d1, 2 haplotypes) in Switzerland dating ~8.0 ky; two (2 haplotypes each) in Ukraine with the oldest subclade (A1a2e1a2a2) dating ~7.6 ky and one (3 haplotypes) in Poland (A1a2e1a1a2a1b) dating ~6.0 ky. This feature is not exclusive to *H. r. rustica*, but it characterizes also *H. r. transitiva*: two subclades (2-3 haplotypes) with the oldest (A1a2e1a3) dating ~11.6 ky (supplementary table S3, Supplementary Material online).

With a lower degree of specificity, some geographic clustering characterizes also a few more common and sometimes older branches. For example, sub-haplogroups A1a1a1a (~19.5 ky), A1a2e1a1a5 (~11.5 ky) and A1a2f1b (~11.1 ky) are over-represented in the Danish population ($\chi^2_{[24]}$ = 10.276, 29.752 and 14.970; P = 0.0028, 0.0001 and 0.0032, respectively) compared to other European locations (supplementary fig. S6, Supplementary Material online).

At the other extreme, we also observed a couple of instances in which specimens sampled at very distant locations harboured the same haplotype (#177 from Denmark and #178 from Italy; #200 from Poland and #201 from Italy). They suggest that long-distance dispersal between populations occurs, in agreement with observations concerning the behavioural flexibility and adaptability of the species (Mead 2002; Turner 2006; Romano et al. 2017; Teglhøj 2020).

While limited by the relatively small size of our population samples and restricted to haplogroup A mitogenomes, the complete or partial clustering of some sub-haplogroups of A would fit with the generally reported short-distance dispersal of offspring from natal to breeding sites, although this feature is less extreme in females compared to males (Balbontín et al. 2009; Scandolara et al. 2014), thus less detectable in terms of mtDNA. On the other hand, the general overall sharing of the haplogroup A branches among *H. r. rustica* populations and between *H. r. rustica* and *H. r. transitiva* can be at least in part explained when considering that even short-distance dispersal can lead to extensive and long-distance gene flow over the course of generations. Moreover, if the instances of long-distance dispersal from natal to breeding sites are confirmed, even at a low percentage, they would further speed up the loss of genetic structure in European barn swallow populations.

As for the remaining three major haplogroups, B, C and D (fig. 1 and supplementary fig. S5, Supplementary Material online), the former encompasses only the four *H. r. transitiva* mitogenomes already mentioned above and is dated at 18.9 ± 3.9 kya. It shares an ancestral node (AB; 115.6 ± 13.3 kya) with the sister haplogroup A, which is approximately 40 ky younger than the CD node (156.4 ± 18.0 kya) from which C and D derive.

Haplogroup C includes only *H. r. gutturalis* samples, four of the five sampled in China and is dated at 31.1 ± 5.7 kya, while the fifth is a member of haplogroup A. The detection of haplogroup A in the *gutturalis* individual might indicate past or present admixture with *rustica*, especially when considering that it was collected in the westernmost (Nujiang Prefecture) of the sampling locations in China, the closest to the breeding range of *H. r. rustica*.

Finally, haplogroup D, dated at 51.1 ± 7.9 kya, characterizes all 15 *H. r. erythrogaster* specimens from North America (USA, Nebraska), in either one or the other of its sub-haplogroups (D1 and D2). Haplogroup D age provides a minimum time for the spread of *H. rustica* from Asia to the Americas, and indicates that North America was most likely the nesting ground of the ancestors of *H. r. erythrogaster* since at least 51 ky ago.

Subspecies Specificity of the Major Haplogroups

To gain a broader view of the haplogroup distribution in the different subspecies, including some not included in our study, we compared the combined *ND2* and *CYB* gene variation of our mitogenomes with that reported in 119 barn swallow mtDNAs available from previous studies (Dor et al. 2010, 2012; Liu et al. 2015, direct submission; Keepers et al. 2016, direct submission; Smith et al. 2018; Feng et al. 2020; Carter et al. 2020) (fig. 3).

The phylogeny of fig. 3 confirms that haplogroup A is typical of both *H. r. rustica* and *H. r. transitiva*, with *H. r. transitiva* mitogenomes scattered among those of *H. r. rustica* in virtually all sub-haplogroups of A. Moreover, it reveals that the four haplogroup B mitogenomes observed in *H. r. transitiva* form a clade that is defined by the transitions at nps 14235 and 14243 in *CYB*. This branch encompasses also an additional *H. r. transitiva* specimen (Dor et al. 2012) and nine of eleven *H. r. savignii* (Dor et al. 2010; Smith et al. 2018). This high frequency of haplogroup B in *H. r. savignii* indicates that haplogroup B is typical of the sedentary Egyptian subspecies. Moreover, the detection of some B mitogenomes in *transitiva* and some A mitogenomes in *savignii* (fig. 3) appears to indicate that gene flow of maternal lineages is not restricted to *transitiva* and *rustica*, but it also occurs between *transitiva* and *savignii*, and possibly also between *rustica* and *savignii*. These and other alternative scenarios cannot be distinguished without nuclear genome data.

As for haplogroup C (n = 10), the phylogeny confirms instead its complete subspecies specificity. It includes only *H. r. gutturalis* specimens: the four from China of this study (#393-396), one from Japan, three from Russia, one from Mongolia (Dor et al. 2010) and one of an undefined Asian origin (Liu et al. 2015, direct submission).

A more complex situation concerns haplogroup D. The phylogeny of fig. 3 supports the exclusive affiliation of all *H. r. erythrogaster* specimens (n = 30) to haplogroup D: the 15 from Nebraska of this study (#397-411), additional 14 from the USA (Dor et al. 2010; Keepers et al. 2016, direct submission; Smith et al. 2018) and one from Argentina (Dor et al. 2010). As already shown by the phylogeny of entire mitogenomes, they all belong to either sub-haplogroups D1 or D2, whose ages are estimated at 19.7 ± 3.9 kya and 20.6 ± 3.4 kya, respectively (supplementary table S3, Supplementary Material online).

However, *ND2* and *CYB* sequences are available also for three *H. r. tytleri* specimens (Dor et al. 2010; Carter et al. 2020), an Asian subspecies that was not included in our survey of entire mitogenomes. The three *H. r. tytleri* partial mtDNA sequences appear to form a private sub-haplogroup within D1, which we named $D1^{tytleri}$ (fig. 3). It is a sister branch to the D1 branches of *H. r. erythrogaster*, thus supporting the previously proposed close relationship between *H. r. tytleri* and *H. r. erythrogaster* as well as an American origin of the ancestral mitogenomes of *H. r. tytleri* (Zink et al. 2006; Dor et al. 2010, 2012). Moreover, taking into consideration that D1 arose approximately 20 kya, we have now a maximum age boundary for the back migration from North America: the ancestors of *H. r. tytleri* did not move to Asia earlier than 20,000 years ago. As for the minimum boundary for this event, it will be defined only by sequencing *H. r. tytleri* mitogenomes.

The Demography of Barn Swallows over Time

The Bayesian skyline plot of fig. 4 shows changes in the effective population size over time for haplogroup A, which is typical of western Eurasia and by far the most represented in our survey encompassing all *H. r. rustica* mitogenomes and 92% of those from *H. r. transitiva*. It underwent two population growth events. The first, very sharp increase occurred ~30 kya, prior to the Last Glacial Maximum (LGM). This was followed by a plateau throughout the LGM and up to the Younger Dryas (YD; 12.9-11.7 kya), when the second increase began, lasting until the end of the Early Holocene Optimum (EHO) ~6 kya (Baker et al. 2017).

Population expansion has been documented during postglacial periods of many other bird populations, in parallel with and thanks to their northward range expansion (Milá et al. 2006; Zink and Gardner 2017). Migratory behavior might have both resulted from and played a role in this

population expansion. Glacial cycles act as switches for the evolutionarily labile migratory behavior. Lacking a suitable habitat, species would retreat to their wintering ranges during glacial maxima, and revert back to long-distance migration during interglacial periods (Zink and Gardner 2017). Our results on haplogroup A mitogenomes are consistent with *H. r. rustica* ancestors expanding northward from the eastern Mediterranean basin, which might have acted as a refugium during the LGM and the Younger Dryas. *H. r. transitiva* would then have mainly derived from specimens/populations that maintained their sedentary behavior, while *H. r. rustica* would descend from those that differentiated and re-acquired a long-distance migratory behavior while expanding northward at the end (~11.7 kya) of the Younger Dryas. These climatic changes, and possibly the increase in energy consumption associated with the re-acquisition of the long-distance migratory behaviour, appear to strongly affect the extent to which selection modulates the evolution of mitochondrial protein-coding genes (supplementary fig. S4, Supplementary Material online). Taking the end of the Younger Dryas (~11.7 kya) as a cut-off in the phylogeny, it is evident that the ratio of divergence at non-synonymous and synonymous sites (dN/dS) is much higher when considering only variants accumulated after the Younger Dryas (0.19 *vs* 0.08, Fisher exact test *p-value* = 0.0001). This is particularly evident for genes encoding subunits of OXPHOS complexes I and V, thus supporting scenarios linking heat production and avian flight ability with mitogenome variation (Shen et al. 2009; Zhong et al. 2020).

Such a scenario would explain the sharing of haplogroup A by *rustica* and *transitiva* and many of its sub-branches, and the "intermingling" of their haplotypes within these clades (fig. 2). However, it would also explain the detection of A sub-haplogroups within localized populations in Europe (*H. r. rustica*). The oldest are in the Mediterranean area, A1a2g in Italy and A1a1b3b in Spain (fig. 2), with ages of 11.6 ky and 11.4 ky, respectively. Thus, they arose shortly after the end of the Younger Dryas. In contrast, the population-specific sub-haplogroups detected further north in Europe arose later with a south to north time profile: A1a2d1 in Switzerland (~8.0 ky), A1a2e1a2a2 in Ukraine (~7.6 ky) and A1a2e1a1a2a1b in Poland (~6.0 ky). Their ages suggest that they arose *in situ* when these different European regions became suitable as nesting grounds. There is also a Levantine counterpart to the European-specific sub-haplogroups. This is represented by A1a2e1a3, the oldest *transitiva*-specific sub-haplogroup, which is dated at ~11.6 kya, again immediately after the Younger Dryas, underscoring its role in the differentiation of *rustica* and *transitiva*.

The chronological gradient from south to north suggests a history of northward expansion from the Near East or the Mediterranean basin at large. This is supported by the negative correlation between nucleotide diversity and latitude that we observed in *H. r. rustica* and *H. r. transitiva* populations, which encompass most of our dataset and were more densely sampled, when they were grouped into the following macro-groups: South (Algeria, Spain, South Italy and Israel), Center (North Italy and Switzerland) and North (Poland, Ukraine and Denmark) (supplementary fig. S7, Supplementary Material online). For haplogroup A1, a correlation (p-value < 0.05) close to 1 was detected, thus confirming the overall reduction of mitogenome variation from South to North, as expected in models envisioning a more recent origin of central and northern European populations.

Recolonization of Europe from refugia following glacial retreat has been documented in a variety of species (Hewitt 1999, 2000; Hansson et al. 2008). The pattern of ever younger population-specific sub-haplogroups suggests a post-glacial expansion without major loss of genetic diversity and supports a relatively slow northward spreading - the so-called "phalanx-model" of colonization, as opposed to a "pioneer model" (Nichols and Hewitt 1994; Excoffier et al. 2009). Such a slow expansion is a feature of species with short dispersal, strict requirements for habitat, and/or dependency on other species (Hewitt 2004). Barn swallows preferentially nest in human structures and are closely associated with human agriculture. Their association with slow-moving agriculturalists might explain the age gradient from south to north Europe that we observed for the population-specific subclades. The first evidence of human built structures dates to around 12-15

kya (Potts 2012), and the expansion of agriculture from the Middle East might have begun as early as 12-13 kya (Salamini et al. 2002; Arranz-Otaegui et al. 2016). Thus, the second population increase is compatible with a role for rising temperatures at the beginning of the Holocene ~12 kya (Taberlet et al. 1998) as well as for the association with human settlements (Smith et al. 2018).

On the Origin of Barn Swallows

Previous comparisons of mtDNA variation in barn swallows, along with that seen in their closest relatives (Dor et al. 2010; Carter et al. 2020), have suggested that the ancestor of all *Hirundo* species most likely originated in Africa, as most of them have African distributions, including *H. aethiopica* and *H. angolensis* that are the closest relatives to *H. rustica* (Carter et al. 2020).

We further assessed this issue by adding the mitogenomes from other *Hirundo* species to the phylogeny of our barn swallows. The combined tree confirms that the closest species are all from Africa (*H. aethiopica, H. angolensis, H. nigrita, H. smithii* and *H. albigularis*), thus supporting Africa as the ancestral continental source of *H. rustica*, and dates the divergence between *H. rustica* and *H. aethiopica*, the closest species, at about 493 kya (supplementary fig. S8, Supplementary Material online). However, it is possible that, when the HrAM arose ~280 kya, the *H. rustica* population from which all modern *H. rustica* mitogenomes derive had already left Africa and entered into Eurasia.

Conclusions

Over the course of years, numerous studies have shown that the information contained in mtDNA is phylogenetically best exploited when the sequence variation of the entire (or almost entire) mitogenome is assessed and the sequencing survey is carried out on numerous specimens sampled throughout the species distribution range. Here we employed this "magnifying glass" approach to reconstruct the genetic history of an iconic species, the barn swallow. Mitogenome data allowed us to build a detailed phylogeny for the species, to determine its coalescence time as well as the ages of its haplogroups, and to better define the matrilineal relationships between subspecies.

According to our data, the female barn swallow carrying HrAM lived 276.9 ± 24.3 kya, which is much earlier than previously thought (Zink et al. 2006; Dor et al. 2010; Scordato et al. 2017). Considering that, due to its reduced effective population size, mtDNA is much more prone to lineage loss and founder events than its autosomal counterpart, an almost 300 ky of mtDNA divergence implies an even older divergence time for most nuclear genes. This allows for plenty of polymorphism in the species complex and, probably, a rather extensive differentiation of its nuclear genome, thus explaining the observed flexibility and adaptability.

In most cases, we observed complete subspecies and geographic specificity of mitogenome haplogroups, arising at different times and different places, which we employed together with estimated radiation times to postulate an overall model for the geographical and temporal spread of barn swallows (fig. 5). According to the mtDNA data, this species left Africa (or its vicinity) almost 300 kya, expanded first into Eurasia and then, at least 51 kya, into the Americas, from where a relatively recent (< 20 kya) back migration to Asia took place. Subspecies differentiation occurred in parallel to the species dispersal, usually much earlier than previously suggested (Smith et al. 2018).

The notable exception to the haplogroup subspecies specificity is represented by the sedentary Levantine *H. r. transitiva* that extensively shares haplogroup A with the migratory *H. r. rustica* and haplogroup B, to a lesser extent, with the Egyptian *H. r. savignii*. We propose that *rustica* and *transitiva* derive from the same population source, which was located in the Levant and had adapted to sedentarism during the LGM. Our data indicate that the two subspecies began to split rather recently, shortly after 11.7 kya at the end of the Younger Dryas. *H. r. rustica* would descend from individuals that re-acquired the long-distance migratory behavior while expanding northward to

regions that were then becoming suitable as nesting grounds. In contrast, *H. r. transitiva* would derive from the Levantine component that remained *in situ* and maintained its sedentary behavior. Since then, however, *transitiva* did not remain genetically isolated, receiving genetic inputs and admixing with migratory *rustica* populations, as shown by the absence of significant correlations between genetic and geographic distances when assessing the shared haplogroups (supplementary fig. S9, Supplementary Material online), as well as the adjacent *savignii*.

This scenario, which is compatible with the presence of some haplogroup B mitogenomes in *transitiva* as well as its behavioural phenotype, is also supported by field and phenotypical observations (Ambrosini et al. 2009; Reiner Brodetzki et al. 2021), including the expression in *H. r. transitiva* of both elongated tail streamers and dark ventral coloration. The first feature is shared with *rustica*, but not its function as a sexual signal, and the second is shared with *savignii* (Vortman et al. 2011). Genetic admixture is also a plausible explanation for the detection of haplogroup A in one of the five *H. r. gutturalis* specimens from China.

Finally, *Hirundo rustica* has been strongly affected by climatic changes in the past. At the beginning of the Holocene its population size began to grow extensively in parallel with temperature increases, and this growth was probably facilitated by the concomitant spread of agriculture and human built structures. It is also evident that climatic changes occurring during the LGM and the Younger Dryas, and the possible resulting changes in migratory behaviour, significantly affected the extent to which selection modulates gene sequence evolution, to a degree that is comparable with that reported in Neolithic animal domestication (Colli et al. 2015). The strong link of this widespread species with climate fluctuations and human activities makes it an excellent indicator for monitoring and assessing the impact of current global changes on wildlife.

Materials and Methods

Samples Analysed for Mitogenome Variation

We completely sequenced a total of 410 barn swallow mitogenomes. An additional *H. r. rustica* sample from Italy (#151) was extrapolated from Formenti et al. (2019). Samples were collected from Europe and Algeria (n = 340; *H. r. rustica*), Israel (n = 50; *H. r. transitiva*), China (n = 5; *H. r. gutturalis*) and the USA (Nebraska) (n = 15; *H. r. erythrogaster*). *H. r. rustica* were sampled in Denmark (n = 31), Italy (n = 171), Poland (n = 35), Spain (n =31), Switzerland (n = 20), Ukraine (n = 20) and Algeria (n = 32), (supplementary fig. S2, Supplementary Material online). We extracted DNA either from tissues (muscle or liver) of specimens found dead (#1, 20, 35, 58, 136, 302) or blood (all remaining samples). We obtained blood samples, under license according to national guidelines and legislation, by capturing individuals with mist-nets at the barns and cowsheds where barn swallows spend the nights during breeding. Venipuncture of the brachial vein, a minimally invasive technique (Arctander 1988), was performed to draw blood. Blood samples were collected and stored either in heparinized glass capillaries or dehydrated in ethanol. Blood samples from Spain and Nebraska arrived in Sodium-EDTA buffer. Detailed information on the barn swallow samples and their mitogenomes is provided in supplementary table S1, Supplementary Material online.

At the time when the analyses were performed, there were nine entire *H. rustica* mitogenomes in GenBank. However, none of these could be included in our analyses of entire mitogenomes due to the following sequence issues: KX398931 (*H. r. erythrogaster*, Keepers et al. 2016, direct submission), a small nuclear mitochondrial (NUMT) sequence in *ND4*; MN356225 (*H. r. erythrogaster*, Feng et al. 2020, lack of *CR1*, *MP*, *ND6* and *ME*; MN843972 (*H. r. tytleri*), MN829439 (*H. r. rustica*), MN830163 (*H. r. savignii*), MN954681 and MN840495 (*H. r. transitiva*), presence of NUMTs in *ND5* (Carter et al. 2020); MN909724 (*H. r. gutturalis*, Thirouin et al. 2020, direct submission), large insertion in RNR2, many gaps throughout the sequence, two large NUMTs in ND3 and ND5; and KP148840 (*H. r. gutturalis*, Liu et al. 2015, direct

submission), numerous NUMTs throughout the sequence. In addition, we extracted 16 low-coverage mitogenomes (eight from *H. r. savignii* and eight from *H. r. erythrogaster)* from the PRJNA323498 BioProject (Smith et al. 2018). They also harboured many gaps. However, most of these samples (n = 22) could be employed in the phylogenetic analyses of *ND2* and *CYB* gene sequences (see below).

DNA Extraction

We obtained genomic DNA extracted from muscle or liver with the ReliaPrep™ (Promega Madison, WI, USA) gDNA Tissue kit, using the standard protocol for animal tissue. We extracted and purified blood samples using phenol-chloroform. These samples were prepared by breaking 1-2 cm of the glass capillary containing the blood (~4 µl) and placing it overnight at 56°C in a 2 ml tube containing: lysis buffer B (400 mM Tris-HCl, pH 8.0; 100 mM EDTA, pH 8.0; 1% SDS), 250 µl of TBS buffer (20 mM Tris-HCl, pH 7.5; 150 mM NaCl), and 40 µl of Proteinase K (20 mg/ml). Samples from Spain and Nebraska were instead extracted with magnetic beads on a 16 Maxwell® RSC 16 instrument using the dedicated Blood DNA Kit (Promega) and employing the "Blood DNA" protocol. Sample preparation in this case was performed by adding 1-2 µl of blood to 300 µl of lysis buffer and 30 µl Proteinase K and incubating overnight at 56°C. Genomic DNAs were eluted into TE buffer (10 mM Tris-Cl, 1 mM EDTA) or elution buffer (Promega).

Sanger Sequencing

The first five barn swallow mitogenomes that we obtained (#1, 20, 35, 136 and 302) were Sanger sequenced. We designed an initial set of oligonucleotide pairs (not shown) using partial sequences of *H. r. rustica* mtDNA genes (mainly *RNR1, RNR2, ND2, CO1* and *CYB*) available in GenBank using the Primer3Plus software (Untergasser et al. 2012) (https://primer3plus.com/primer3web/primer3web_input.htm). Then, by using a primer walking approach, we designed a set of oligonucleotide pairs that allowed the amplification of the entire mitogenome in eleven overlapping PCR fragments (supplementary table S5, Supplementary Material online) and a set of 33 additional oligonucleotides for sequencing (supplementary table S6, Supplementary Material online). This allowed the amplification and sequencing of mitogenome #20, our first *H. r. rustica* complete mitogenome. The other four mitogenomes (#1, 35, 136 and 302) were then obtained by carrying out PCR reactions with a standard reaction mix (25 µl) containing 1X Buffer (1.5 mM MgCl2), 0.2 mM of each dNTP, 0.6 U of GoTaq G2 Polymerase (Promega), 0.2 µM of each primer and 20–30 ng of DNA, using the following PCR conditions: 94°C (2 min); 35 cycles of 94°C (30 s), 55°C (30 s), 72°C (2 min); and a final extension of 72°C (10 min). PCR products were visualized on a 1% agarose gel and amplicons were sequenced with standard dideoxy sequencing using BigDye v3.1 Chemistry (Applied Biosystems) on 3730xl and 3130xl Genetic Analyzer (Applied Biosystems) following the manufacturer's protocol. Mitogenome #20 (MZ905359) was employed as the *Hirundo rustica* reference sequence (HrrRS). We then obtained the locus map of the barn swallow (supplementary fig. S1, Supplementary Material online) using the CGview Server (Grant and Stothard 2008).

NGS Sequencing

We obtained 405 additional barn swallow mitogenomes by NGS sequencing and extrapolated one more from Formenti et al. (2019). We designed a set of three oligonucleotide pairs with similar T_m (~60°C) and length (20 nt) (supplementary table S7, Supplementary Material online) to amplify the entire mtDNA molecule in three overlapping, long range PCR fragments of comparable lengths (~6400 bps). Each fragment overlapped the next one by about 300-500 bps. PCRs were carried out in 50 µl reaction mix containing 1x GoTaq® Long PCR Master Mix (Promega), 0.3 µM of each primer and ~150 ng of DNA template, using the following 2 step PCR thermal profile: 94°C (2 min); 20 cycles at 94°C (30 s), 58°C (30 s), 65°C (7 min), followed by 10 cycles at 94°C (30 s), 55°C (30 s), 65°C (7 min) and a final extension at 72°C (10 min). PCR products were checked by electrophoresis on 1% agarose gels. PCR purification was performed with the membrane method,

Presto™ 96 Well PCR Cleanup Kit (Geneaid). Amplicons were quantified with the Quantus™ fluorometer (Promega) using the QuantiFluor® ONE dsDNA system. 1 ng for each of the three amplicons were combined for library preparation. The sequencing library was prepared with the Nextera™ DNA Flex Library Prep, following the manufacturer's protocol steps: tagmentation of input DNA, amplification of tagmented DNA with the addition of pre-mixed dual-indexed adapters (IDT® for Illumina Nextera UD indexes or Nextera™ DNA CD Indexes) and PCR clean-up. Libraries were then checked on a 2% agarose gel, quantified using the Quantus™ fluorometer (Promega), normalized and pooled together. We then run the pooled normalized library on the 4150 TapeStation System (Agilent) and diluted to 4 nM using RSB resuspension buffer. Five µl of pooled libraries were denatured using 5 µl of freshly prepared NaOH (0.2N) and diluted to the loading concentration of 6 pM (600 µl final volume) using HT1 hybridization buffer. This was finally sequenced on a MiSeq system (Illumina) using paired-end sequencing either with the MiSeq Reagent Kit v2, (2 x 150 or 2 x 250 cycles) or the MiSeq Reagent Kit v2 Nano (2 x 150 cycles). We also NGS sequenced mitogenome #20 (HrrRS) and the other four samples already Sanger sequenced (#1, 35, 136 and 302), fully confirming the initial Sanger sequences.

Analysis of Mitogenome Sequence Data

The raw MiSeq sequencer data were output in BCL format, demultiplexed and converted to FASTQ format with the Illumina® bash package, bcl2fastq2 Conversion Software v2.20, and trimmed with Trim Galore! v 0.6.4 (Krueger 2012) to remove low quality reads and adapters. We checked the quality of paired end reads by using FastQC v 0.11.9 (Andrews 2010). Files were subsequently converted from FASTQ to BAM by aligning and mapping the reads to the mitogenome #20 (HrrRS) using BWA v0.7.17 (Li 2013, direct submission). BAM files were analysed with Geneious 8.1.5 (Biomatters, Kearse et al. 2012). Variant calling was performed by setting the threshold for heteroplasmies at 30% of reads and considering only mutations in the range 30%-70% as heteroplasmic in phylogenetic analyses. Mitogenomes employed in the phylogenetic analyses were completely sequenced with an average depth of 459X. Finally, samples were exported in the standard FASTA format.

Phylogenetic Analyses and Age Estimates of MtDNA Haplogroups

We built a maximum parsimony (MP) tree by manually programming the software mtPhyl v 5.003 (Eltsov and Volodko 2014) for the analysis of barn swallow mitogenomes (the modified .txt files are available on request). By comparing the mtDNA FASTA sequences to the HrrRS, the software allowed for the reconstruction of a coding-region (15601 bps; nps 1-14859, nps 16068-16740, nps 18075-18143) MP tree with detailed information concerning mutations (supplementary fig. S5, Supplementary Material online). We did not consider indels for tree construction. The tree was rooted using the available *Hirundo angolensis* (NC_050287) and *Hirundo aethiopica* (NC_050293) reference mitogenomes and its topology was also confirmed by using MEGAX (figs. 1 and 2) (Kumar et al. 2018).

We estimated ML and Bayesian ages of haplogroups and sub-haplogroups by using all barn swallow (n = 411) coding-region sequences (15601 bps). We performed ML estimations using the BaseML package present in the PAMLX 1.3.1 software (Xu and Yang 2013) assuming the HKY85 (Hasegawa et al. 1985) mutation model with gamma-distributed (32 categories) rates (plus invariant sites) and 17 partitions (13 for protein coding genes, 1 for tRNAs, 1 for each rRNA gene, and 1 for intergenic regions) using the predefined tree obtained by the MP approach. We converted ML mutational distances into years by assuming an estimated split time between *Progne* (not shown) and *Hirundo* at 9.34 Mya (95% CI: 5.8-13.2 Mya) (Moyle et al. 2016), thus employing the standard approach that does not include the error on the calibration point.

We performed Bayesian estimations using BEAST 2.6.0 (Bouckaert et al. 2019) under the HKY substitution model (gamma-distributed rates plus invariant sites) with a relaxed clock (log normal). We entered as prior the clock value of 2.45 $x10^{-8}$ base substitution per nucleotide per year (or one

mutation every 2616 years), derived from the rate calculated in the ML method. The chain length was established at 100,000,000 iterations, with samples drawn every 1,000 Markov chain Monte Carlo (MCMC) steps after a discarded burn-in of 10,000,000 steps (Olivieri et al. 2017). We performed the demographic analysis on the BEAST results using Tracer v1.7.1 (Rambaut et al. 2018) and Excel using a generation time of one year. We constructed Bayesian skyline plots (BSPs) using the output tree file and a stepwise constant.

To assess the subspecies specificity of major haplogroups in a wider sample encompassing also previously published sequences, we built a MP tree based on *ND2* (1017 bps; nps 3980-4996) and *CYB* (1058 bps; nps 13696-14753) gene sequences (fig. 3), for a total of 2075 bps. We aligned sequences with MEGAX and rooted the tree using the *H. aethiopica* (NC_050293) and *H. angolensis* (NC_050287) reference mitogenomes.

Mitogenome Diversity and Latitude

To identify a potential correlation between mitogenome diversity (entire mitogenomes) and latitude, we measured both haplotype diversity (HD) and nucleotide diversity (Pi) for the most represented haplogroups (A1 as a whole as well as its major sub-haplogroups A1a1 and A1a2) using DNAsp 6.12.03 (Rozas et al. 2017). Both indices were compared with the average of the latitudes among the samples of each population using the software Tableau 2021.4 (https://www.tableau.com/).

Isolation by Distance

To assess Isolation by Distance (IBD), genetic distance matrices were created based on PhiSt, which was computed using the R package *haplotypes* (Aktas 2020), while pairwise geographic distances were calculated with the *geodist* R package (Padgham and Sumner 2021), applying the geodesic methods provided in (Karney 2013). We tested the correlations between genetic and geographic matrices for the most represented haplogroups (A1 as a whole as well as its major sub-haplogroups A1a1 and A1a2) using a Mantel test and simple linear regression, to also account for possible type II errors (Teske et al. 2018). The Mantel test was performed using the R package *ade4* (Dray and Dufour 2007) with 999 permutations. This program tests for significant IBD by comparing the observed correlation with a histogram of simulated correlation categories and their frequency under the assumption of no IBD. Simple linear regression was performed in the *stats* package (R Core Team 2021). We generated plots in R using the package *ggplot2* (Wickham 2016).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgements

We dedicate this study to the late Nicola Saino, who greatly contributed to its initial conception, planning and sample collection. We thank Dan Liang and Yanyan Zhao for sampling. This research received support from the Italian Ministry of Education, University and Research (MIUR): the project PRIN2017 2017CWHLHY (A.T. and L.G.) and Dipartimenti di Eccellenza Program (2018–2022) - Department of Biology and Biotechnology "L. Spallanzani" University of Pavia (to A.O., O.S., A.A., L.F. and A.T.); Junta de Extremadura / FEDER reference IB20089 (to A.M.); the "Fondazione Adriano Buzzati – Traverso" for the L. Luca Cavalli-Sforza fellowship (to N.R.M).

Author Contributions

A.T., G.F. and L.G. conceived the project; A.T., A.O. and L.F. planned the experimental approach and supervised experiments/computational analyses and data; G.L. performed experiments and computational analyses with contributions by G.F., E.M., L.C., H.L., S.S., G.R.G., M.R.C., N.R.M., G.C. and M.G; A.C., A.M., N.S., M.C., A.R.A., R.A., D.R., L.G., Y.L., Y.K., A.M., C.L-C., J.B.A., T.A.M., P.M. and A.P.M. performed/contributed to the sample collection; C.C., V.C., O.S. and A.A. provided reagents/materials/analytical resources; G.L., L.F., A.O. and A.T. wrote the manuscript; all authors reviewed and approved the manuscript.

Data Availability

The sequence data for the 411 *H. rustica* mitogenomes are available in GenBank under accession numbers MZ905359, OK539050-OK539458 and OK624420. The raw NGS sequence data (fastq and bam files) are available under the ENA accession number PRJEB51610.

References

Achilli A, Olivieri A, Pellecchia M, Uboldi C, Colli L, Al-Zahery N, Accetturo M, Pala M, Hooshiar Kashani B, Perego UA, et al. 2008. Mitochondrial genomes of extinct aurochs survive in domestic cattle. Curr Biol. 18(4):R157-158.

Achilli A, Olivieri A, Soares P, Lancioni H, Hooshiar Kashani B, Perego UA, Nergadze SG, Carossa V, Santagostino M, Capomaccio S, et al. 2012. Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proc Natl Acad Sci USA. 109(7):2449-2454.

Aktas C. 2020. haplotypes: Manipulating DNA Sequences and Estimating Unambiguous Haplotype Network with Statistical Parsimony. R package version 1.1.2. https://CRAN.R-project.org/package=haplotypes

Ambrosini R, Møller AP, Saino N. 2009. A quantitative measure of migratory connectivity. J Theor Biol. 257(2):203-211.

Ambrosini R, Rubolini D, Trovò P, Liberini G, Bandini M, Romano A, Sicurella B, Scandolara C, Romano M, Saino N. 2012. Maintenance of livestock farming may buffer population decline of the Barn Swallow *Hirundo rustica*. Bird Conserv Int. 22:411–428.

Andrews S. 2010. FastQC a quality control tool for high throughput sequence data. Babraham bioinformatics. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arctander P. 1988. Comparative studies of avian DNA by restriction fragment length polymorphism analysis: convenient procedures based on blood samples from live birds. J Ornithol. 129:205-216.

Areta JI, Salvador SA, Gandoy FA, Bridge ES, Gorleri FC, Pegan TM, Gulson-Castillo ER, Hobson KA, Winkler DW. 2021. Rapid adjustments of migration and life history in hemisphere-switching cliff swallows. Curr Biol. 31(13):2914-2919.e2.

Arranz-Otaegui A, Colledge S, Zapata L, Teira-Mayolini LC, Ibáñez JJ. 2016. Regional diversity on the timing for the initial appearance of cereal cultivation and domestication in southwest Asia. Proc Natl Acad Sci USA. 113(49):14001-14006.

Baker JL, Lachniet MS, Chervyatsova O, Asmerom Y, Polyak VJ. 2017. Holocene warming in western continental Eurasia driven by glacial retreat and greenhouse forcing. Nat Geosci. 10: 430–435.

Balbontín J, Møller AP, Hermosell IG, Marzal A, Reviriego M, de Lope F. 2009. Geographic patterns of natal dispersal in barn swallows *Hirundo rustica* from Denmark and Spain. Behav Ecol Sociobiol. 63:1197–1205.

Barth JMI, Damerau M, Matschiner M, Jentoft S, Hanel R. 2017. Genomic differentiation and demographic histories of Atlantic and Indo-Pacific yellowfin tuna (*Thunnus albacares*) populations. Genome Biol Evol. 9(4):1084-1098.

Battaglia V, Gabrieli P, Brandini S, Capodiferro MR, Javier PA, Chen XG, Achilli A, Semino O, Gomulski LM, Malacrida AR, et al. 2016. The worldwide spread of the tiger mosquito as revealed by mitogenome haplogroup diversity. Front Genet. 7:208.

Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, Kivisild T, Torroni A, Villems R. 2012. "A Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet. 90(4):675-684.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, et al. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comp Biol. 15(4):e1006650.

Brown CR, Brown MB. 1999. Barn Swallow (*Hirundo rustica*). The birds of North America. Inc. Philadelphia. 452:32.

Carter JK, Innes P, Goebl AM, Johnson B, Gebert M, Attia Z, Gabani Z, Li R, Melie T, Dart C, et al. 2020. Complete mitochondrial genomes provide current refined phylogenomic hypotheses for relationships among ten *Hirundo* species. Mitochondrial DNA B Resour. 5(3):2881-2885.

Cheng T. 1987. A Synopsis to the Avifauna of China. Science Press, Paul Parey, Hamburg, 1987. ISBN 10: 3490125185.
Cole TL, Ksepka DT, Mitchell KJ, Tennyson AJD, Thomas DB, Pan H, Zhang G, Rawlence NJ, Wood JR, Bover P, et al. 2019. Mitogenomes uncover extinct penguin taxa and reveal island formation as a key driver of speciation. Mol Biol Evol. 36(4):784-797.
Colli L, Lancioni H, Cardinali I, Olivieri A, Capodiferro MR, Pellecchia M, Rzepus M, Zamani W, Naderi S, Gandini F, et al. 2015. Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability. BMC Genomics. 16:1115.
de Manuel M, Barnett R, Sandoval-Velasco M, Yamaguchi N, Garrett Vieira F, Zepeda Mendoza ML, Liu S, Martin MD, Sinding MS, Mak SST et al. 2020. The evolutionary history of extinct and living lions. Proc Natl Acad Sci USA. 117(20):10927-10934.
Dickinson EC, Dekker RWRJ. 2001. Systematic notes on Asian birds. 13. A preliminary review of the Hirundinidae. Zool Verh. 335:127-144.
Dickinson EC, Eck S, Milensky CM. 2002. Systematic notes on Asian birds. 31. Eastern races of the barn swallow *Hirundo rustica* Linnaeus, 1758. Zool Verh. 340:201-203.
Dor R, Safran RJ, Sheldon FH, Winkler DW, Lovette IJ. 2010. Phylogeny of the genus *Hirundo* and the barn swallow subspecies complex. Mol Phylogenet Evol. 56(1):409-418.
Dor R, Safran RJ, Vortman Y, Lotem A, McGowan A, Evans MR, Lovette IJ. 2012. Population genetics and morphological comparisons of migratory European (*Hirundo rustica rustica*) and sedentary East-Mediterranean (*Hirundo rustica transitiva*) barn swallows. J Hered. 103(1):55-63.
Dray S, Dufour A. 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. J Stat Softw. 22(4):1–20.
Eltsov N, Volodko N. 2014. mtPhyl-software tool for human mtDNA analysis and phylogeny reconstruction. Available from: http:/eltsov.org.
Excoffier L, Foll M, Petit RJ. 2009. Genetic consequences of range expansions. Annu Rev Ecol Evol Syst. 40:481–501.
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(7833):252-257. Erratum in: Nature. 2021; 592(7856):E24.
Formenti G, Chiara M, Poveda L, Francoijs KJ, Bonisoli-Alquati A, Canova L, Gianfranceschi L, Horner DS, 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(1):giy142.
Grant JR, Stothard P. 2008. The CGView Server: a comparative genomics tool for circular genomes. Nucleic Acids Res. 36 (Web Server issue):W181-184.
Green A. 1988. Cultural responses to the migration of the barn swallow in Europe, in Stanley Cramp (ed.), Handbook of the Birds of Europe, the Middle East and North Africa: The Birds of the Western Palearctic, vol. 5 (Oxford & New York: Oxford University Press, 1988), 263ff.
Hansson B, Hasselquist D, Tarka M, Zehtindjiev P, Bensch S. 2008. Postglacial colonisation patterns and the role of isolation and expansion in driving diversification in a passerine bird. Plos One. 3:e2794.
Hewitt GM. 1999. Post-glacial re-colonization of European biota. Biol J Linn Soc. 68:87–112.
Hewitt GM. 2000. The genetic legacy of the Quaternary ice ages. Nature. 405:907–913.
Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. Phil Trans R Soc B Biol Sci. 359:183–195.
Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22(2):160-174.

Hobson KA, Kardynal KJ, Van Wilgenburg SL, Albrecht G, Salvadori A, Cadman MD, Liechti F, Fox JW. 2015. A continent-wide migratory divide in North American breeding barn swallows (*Hirundo rustica*). PLoS One. 10(6):e0129340.
Karney CFF. 2013. Algorithms for geodesics. J Geod. 87:43-55.
Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28(12):1647-1649.
Keepers KG, Scordato ESC, Jenkins B, Safran RJ, Kane NC. 2016. The complete annotated mitochondrial genome of the North American barn swallow, *Hirundo rustica erythrogaster*. Direct submission.
Kiat Y, Izhaki I, Sapir N. 2019. The effects of long-distance migration on the evolution of moult strategies in Western-Palearctic passerines. Biol Rev. 94(2):700-720.
Krueger F. 2012. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. Available from: http://www.bioinformatics.babraham. ac.uk/projects/trim_galore/.
Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol Biol Evol. 35(6):1547-1549.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013/05/26. https://arxiv.org/abs/1303.3997v2.
Liechti F, Scandolara C, Rubolini D, Ambrosini R, Korner-Nievergelt F, Hahn S, Lardelli R, Romano M, Caprioli M, Romano A, Sicurella B. 2015. Timing of migration and residence areas during the non-breeding period of barn swallows *Hirundo rustica* in relation to sex and population. J Avian Biol. 46(3):254-265.
Liu J, Liu S, Shao C, Zhang Y, Xie Y, Tang Q, Shen Y. Xie J. 2015. Characterization of the complete mitochondrial genome of *Hirundo rustica.* Direct submission.
Liu Y, Scordato ES, Zhang Z, Evans M, Safran RJ. 2020. Analysing phenotypic variation in barn swallows (*Hirundo rustica*) across China to assess subspecies status. Biol J Linn Soc. 131(2): 319-331.
Mackiewicz P, Urantówka AD, Kroczak A, Mackiewicz D. 2019. Resolving phylogenetic relationships within passeriformes based on mitochondrial genes and inferring the evolution of their mitogenomes in terms of duplications. Genome Biol Evol. 11(10):2824-2849.
Malaitad T, Laipasu P, Eiamampai K, Poeaim S. 2016. Identification of the subspecies and gender of barn swallow (*Hirundo rustica*). Int J Agric Technol. 12.7.1:1549-1556.
Mead C. 2002. Barn Swallow *Hirundo rustica*. In The Migration Atlas: movements of the birds of Britain and Ireland. London, T. & A.D. Poyser.
Miao YW, Peng MS, Wu GS, Ouyang YN, Yang ZY, Yu N, Liang JP, Pianchou G, Beja-Pereira A, Mitra B, et al. 2013. Chicken domestication: an updated perspective based on mitochondrial genomes. Heredity (Edinb). 110(3):277-282.
Milá B, Smith TB, Wayne RK. 2006. Postglacial population expansion drives the evolution of long-distance migration in a songbird. Evolution 60:2403–2409.
Møller AP. 1994. Sexual selection and the barn swallow. Oxford University Press.
Møller AP. 2019. Parallel declines in abundance of insects and insectivorous birds in Denmark over 22 years. Ecol Evol. 9(11):6581-6587.
Morin PA, Parsons KM, Archer FI, Ávila-Arcos MC, Barrett-Lennard LG, Dalla Rosa L, Duchêne S, Durban JW, Ellis GM, Ferguson SH, et al. 2015. Geographic and temporal dynamics of a global radiation and diversification in the killer whale. Mol Ecol. 24(15):3964-3979.
Moyle RG, Oliveros CH, Andersen MJ, Hosner PA, Benz BW, Manthey JD, Travers SL, Brown RM, Faircloth BC. 2016. Tectonic collision and uplift of Wallacea triggered the global songbird radiation. Nat Commun. 7:12709.

Nichols RA, Hewitt GM. 1994. The genetic consequences of long-distance dispersal during colonization. Heredity. 72:312–317.
Niedziałkowska M, Tarnowska E, Ligmanowska J, Jędrzejewska B, Podgórski T, Radziszewska A, Ratajczyk I, Kusza S, Bunevich AN, Danila G, et al. 2021 Clear phylogeographic pattern and genetic structure of wild boar *Sus scrofa* population in Central and Eastern Europe. Sci Rep. 11(1):9680.
Olivieri A, Sidore C, Achilli A, Angius A, Posth C, Furtwängler A, Brandini S, Capodiferro MR, Gandini F, Zoledziewska M, et al. 2017. Mitogenome diversity in Sardinians: A genetic window onto an island's past. Mol Biol Evol. 34(5):1230-1239.
Padgham M, Sumner MD. 2021. geodist: Fast, Dependency-Free Geodesic Distance Calculations. R package version 0.0.7. https://CRAN.R-project.org/package=geodist
Peng MS, Xu W, Song JJ, Chen X, Sulaiman X, Cai L, Liu HQ, Wu SF, Gao Y, Abdulloevich NT, et al. 2018. Mitochondrial genomes uncover the maternal history of the Pamir populations. Eur J Hum Genet. 26(1):124-136.
Potts R. 2012. Evolution and environmental change in early human prehistory. Annu Rev Anthropol. 41:151-167.
Quinn TW, Wilson AC. 1993. Sequence evolution in and around the mitochondrial control region in birds. J Mol Evol. 37(4):417-425.
R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 67(5):901-904.
Reiner Brodetzki T, Lotem A, Safran RJ, Hauber ME. 2021. Lack of subspecies-recognition in breeding barn swallows (*Hirundo rustica transitiva*). Behav Processes. 189:104422.
Romano A, Saino N, Møller AP. 2017. Viability and expression of sexual ornaments in the barn swallow *Hirundo rustica*: a meta-analysis. J Evol Biol. 30(10):1929-1935.
Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 34(12):3299-3302.
Safran RJ, Scordato ES, 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(16):3865-3883.
Salamini F, Ozkan H, Brandolini A, Schäfer-Pregl R, Martin W. 2002. Genetics and geography of wild cereal domestication in the Near East. Nat Rev Genet. 3(6):429-441.
Santure AW, Ewen JG, Sicard D, Roff DA, Møller AP. 2010. Population structure in the barn swallow, *Hirundo rustica*: a comparison between neutral DNA markers and quantitative traits. Biol J Linn Soc. 99(2):306-314.
Scandolara C, Lardelli R, Sgarbi G, Caprioli M, Ambrosini R, Rubolini D, Saino N. 2014. Context-, phenotype-, and kin-dependent natal dispersal of barn swallows (*Hirundo rustica*). Behav Ecol. 25(1):180–190.
Scordato ESC, Smith CCR, Semenov GA, Liu Y, Wilkins MR, Liang W, Rubtsov A, Sundev G, Koyama K, Turbek SP, et al. 2020. Migratory divides coincide with reproductive barriers across replicated avian hybrid zones above the Tibetan Plateau. Ecol Lett. 23(2):231-241.
Scordato ESC, Wilkins MR, Semenov G, Rubtsov AS, Kane NC, Safran RJ. 2017. Genomic variation across two barn swallow hybrid zones reveals traits associated with divergence in sympatry and allopatry. Mol Ecol. 26(20):5676-5691.
Sheldon FH, Whittingham LA, Moyle RG, Slikas B, Winkler DW. 2005. Phylogeny of swallows (Aves: Hirundinidae) estimated from nuclear and mitochondrial DNA sequences. Mol Phylogenet Evol. 35(1):254-270.

Shen YY, Shi P, Sun YB, Zhang YP. 2009. Relaxation of selective constraints on avian mitochondrial DNA following the degeneration of flight ability. Genome Res. 19(10):1760-1765.
Shirihai H, Dovrat E, Christie DA, Harris A. 1996. The birds of Israel. London, Academic Press.
Smith CCR, Flaxman SM, Scordato ESC, Kane NC, Hund AK, Sheta BM, Safran RJ. 2018. Demographic inference in barn swallows using whole-genome data shows signal for bottleneck and subspecies differentiation during the Holocene. Mol Ecol. 27(21):4200-4212.
Stewart JB, Freyer C, Elson JL, Wredenberg A, Cansu Z, Trifunovic A, Larsson NG. 2008. Strong purifying selection in transmission of mammalian mitochondrial DNA. PLoS Biol. 6(1):e10.
Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF. 1998. Comparative phylogeography and postglacial colonization routes in Europe. Mol Ecol. 7(4):453-464.
Teglhøj PG. 2020. Natal dispersal and recruitment of Barn Swallows *Hirundo rustica* in an urban habitat. Bird Study. 67(4):420-428.
Teske PR, Golla TR, Sandoval-Castillo J, Emami-Khoyi A, van der Lingen CD, von der Heyden S, Chiazzari B, Jansen van Vuuren B, Beheregaray LB. 2018. Mitochondrial DNA is unsuitable to test for isolation by distance. Sci Rep. 8(1):8448.
Thirouin KR, Goebl AM, Carter J, Scordato E, Hund A, Safran RJ, KaneN. Ecol Evol Biol. 2020. Direct submission.
Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. 2006. Harvesting the fruit of the human mtDNA tree. Trends Genet. 22(6):339-345.
Turner A. 2006. The Barn Swallow. London, T. & A.D. Poyser.
Turner A, Rose C. 2010. A handbook to the swallows and martins of the world. London, A & C Black.
Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3--new capabilities and interfaces. Nucleic Acids Res. 40(15):e115.
Urantówka AD, Kroczak A, Mackiewicz P. 2020. New view on the organization and evolution of Palaeognathae mitogenomes poses the question on the ancestral gene rearrangement in Aves. BMC Genomics. 21(1):874.
von Rönn JA, Shafer AB, Wolf JB. 2016. Disruptive selection without genome-wide evolution across a migratory divide. Mol Ecol. 25(11):2529-2541.
Vortman Y, Lotem A, Dor R, Lovette IJ, Safran RJ. 2011. The sexual signals of the East-Mediterranean barn swallow: a different swallow tale. Behavioral Ecol. 22(6):1344-1352.
Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
Winkler DW, Gandoy FA, Areta JI, Iliff MJ, Rakhimberdiev E, Kardynal KJ, Hobson KA. 2017. Long-distance range expansion and rapid adjustment of migration in a newly established population of barn swallows breeding in Argentina. Curr Biol. 27(7):1080-1084.
Xu B, Yang Z. 2013. PAMLX: a graphical user interface for PAML. Mol Biol Evol. 30(12):2723-2724.
Zhong Y, Zhou M, Ouyang B, Zeng C, Zhang M, Yang J. 2020. Complete mtDNA genome of *Otus sunia* (Aves, Strigidae) and the relaxation of selective constrains on Strigiformes mtDNA following evolution. Genomics. 112(5):3815-3825.
Zink RM, Gardner AS. 2017. Glaciation as a migratory switch. Sci Adv. 3:e1603133.
Zink RM, Pavlova A, Rohwer S, Drovetski SV. 2006. Barn swallows before barns: population histories and intercontinental colonization. Proc Biol Sci. 273(1591):1245-1251.

Figure Legends

Fig. 1. Schematic maximum parsimony phylogeny of *Hirundo rustica* mitogenomes. This tree was built using the entire mitogenome coding-region (15601 bps; nps 1-14859, nps 16068-16740, nps 18075-18143) of 411 barn swallows. It was rooted using *H. angolensis* (NC_050287) and *H. aethiopica* (NC_050293) reference mitogenomes (not displayed). Haplogroups are represented as triangles whose bases are proportional to the number of mitogenomes. HrAM refers to the *Hirundo rustica* Ancestral Mitogenome. Different colours were assigned only to major branches. Bootstrap values (1000 iterations) are shown. The timeline (log_{10}) at the bottom refers to the Bayesian coalescence times of supplementary table S3, Supplementary Material online. The insets A and B illustrate the frequencies of the major haplogroups in the different sampling locations. Inset B details frequencies in Italy and Switzerland.

Fig. 2. Schematic maximum parsimony phylogeny of haplogroup A mitogenomes. This tree was built using the coding regions of 388 mitogenomes belonging to haplogroup A (fig. 1) and was rooted with the available *H. angolensis* (NC_050287) and *H. aethiopica* (NC_050293) reference mitogenomes (not displayed). Mitogenomes marked in black are from *H. r. transitiva* specimens sampled in Israel, the one in white (see also red arrow) is from a *H. r. gutturalis* sample (#258) from China, while all others are from *H. r. rustica*. Main haplogroup affiliations are shown, with branches coloured according to fig. 1. Branch lengths are proportional to the number of nucleotide substitutions. Bootstrap values (1000 iterations) are shown only for the deepest nodes. Six country-specific sub-haplogroups are also shown. They are the oldest found in the reported country. Additional details about samples and mitogenomes are provided in supplementary table S1, Supplementary Material online.

Fig. 3. Maximum parsimony phylogeny of *Hirundo rustica ND2* and *CYB* gene sequences. This tree includes 155 barn swallows from different subspecies for which both *ND2* and *CYB* gene sequences were available. A total of 119 are from the literature and the remaining were selected from our mitogenome dataset as follows: the first five mitogenomes that we obtained from *H. r. rustica* (#1, 20, 35, 136, 302) and the one from Formenti et al. (2019) (#151), all mitogenomes from the uncommon sub-haplogroups of A (A1b, A2, A3; #383-388) and all mitogenomes belonging to haplogroups B, C and D (#389-411) (supplementary table S1, Supplementary Material online). Sequences encompass 2075 bps, 1017 bps of *ND2* (nps 3980-4996) and 1058 bps of *CYB* (nps 13696-14753). The tree was rooted using the *H. aethiopica* and *H. angolensis* reference mitogenomes. Main haplogroup and sub-haplogroup affiliations are shown. Colours identify the different subspecies. The numbers on the branches indicate the number of distinguishing mutations while the numbers in parentheses refer to the following publication sources: (1) this study; (2) Carter et al. (2020); (3) Dor et al. (2010); (4) Dor et al. (2012); (5) Liu et al. (2015), direct submission; (6) Keepers et al. (2016), direct submission; (7) Smith et al. (2018); (8) Feng et al. (2020). Sequences not covering the aforementioned *ND2* and *CYB* gene ranges were not included, as well as sequences that harboured gaps at informative nucleotides. The two mtDNAs forming the rather long sub-branch (6 mutations) within A1a2, one from *H. r. savignii* and one from *H. r. transitiva* (Carter et al. 2020), most likely contain erroneous mutations as their mitogenome sequences harboured NUMTs (see Materials and Methods). A similar problem characterizes the *H. r. gutturalis* sequence KP148840 (Liu et al. 2015, direct submission) with its 81 mutations branch.

Fig. 4. Bayesian skyline plot (BSP) of haplogroup A mitogenomes. The plot considers the 388 haplogroup A mitogenomes listed in supplementary table S1, Supplementary Material online. These include all *H. r. rustica* mitogenomes and 92% of those from *H. r. transitiva*. The black line indicates the median estimate of the effective population size and the blue shading shows the 95% highest posterior density limits. The time axis is limited to 50 kya, beyond which the curve remains flat. EHO; Early Holocene Optimum, YD, Younger Dryas; LGM, Last Glacial Maximum.

Fig. 5. A model for the geographical and temporal spread of barn swallows. Map showing time divergence and hypothetical splits and diffusion routes of barn swallow haplogroups prior to the Younger Dryas and the subsequent climatic changes. The dashed grey circle indicates the possible homeland of the *Hirundo rustica* ancestral mitogenome (HrAM), while the other dashed circles indicate zones where two haplogroups are currently found, possibly indicating recent admixture between subspecies. Colours indicate the breeding ranges of the eight postulated barn swallow subspecies, while striped areas indicate wintering ranges (modified from Turner 2006).

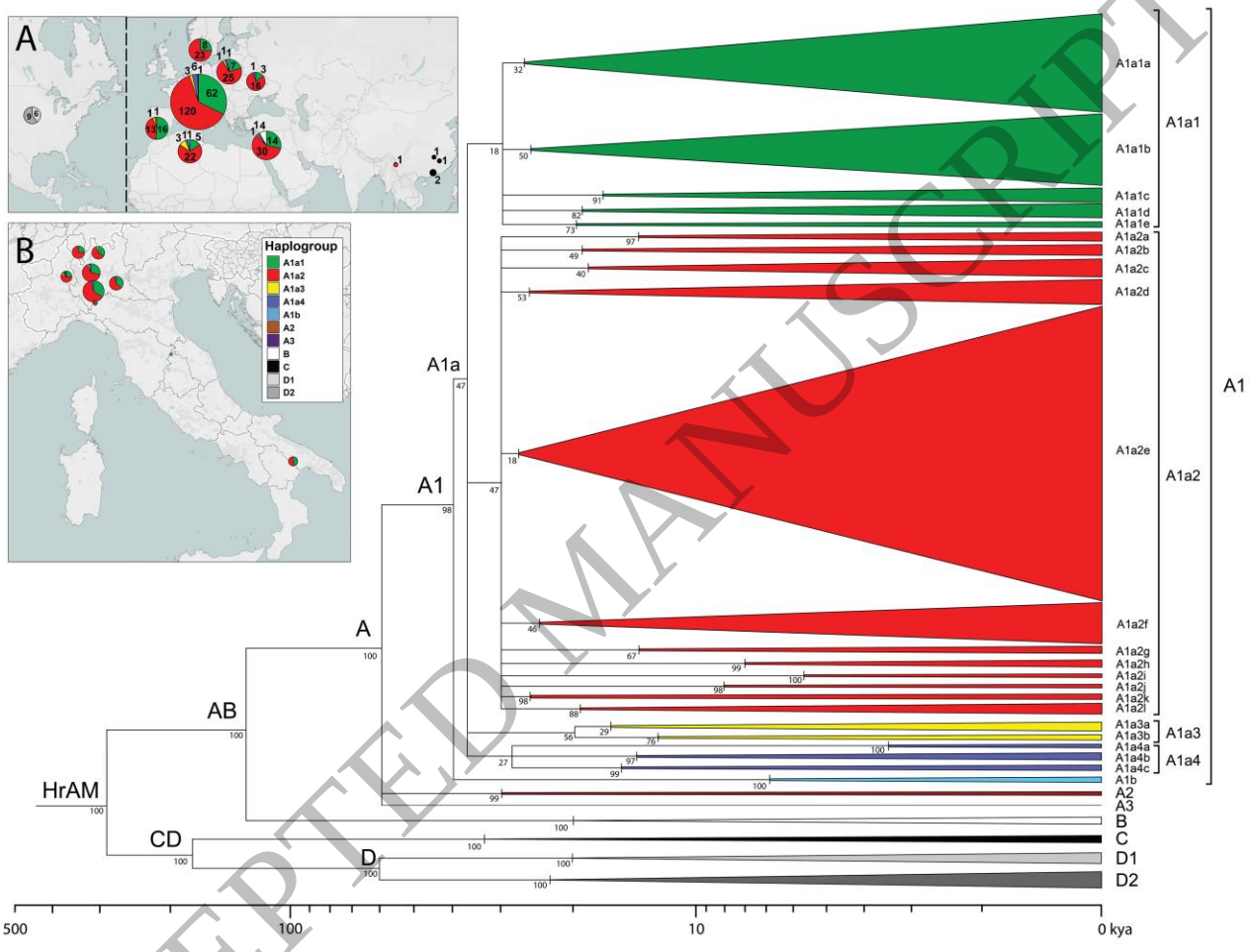


Figure 1
170x128 mm (.76 x DPI)

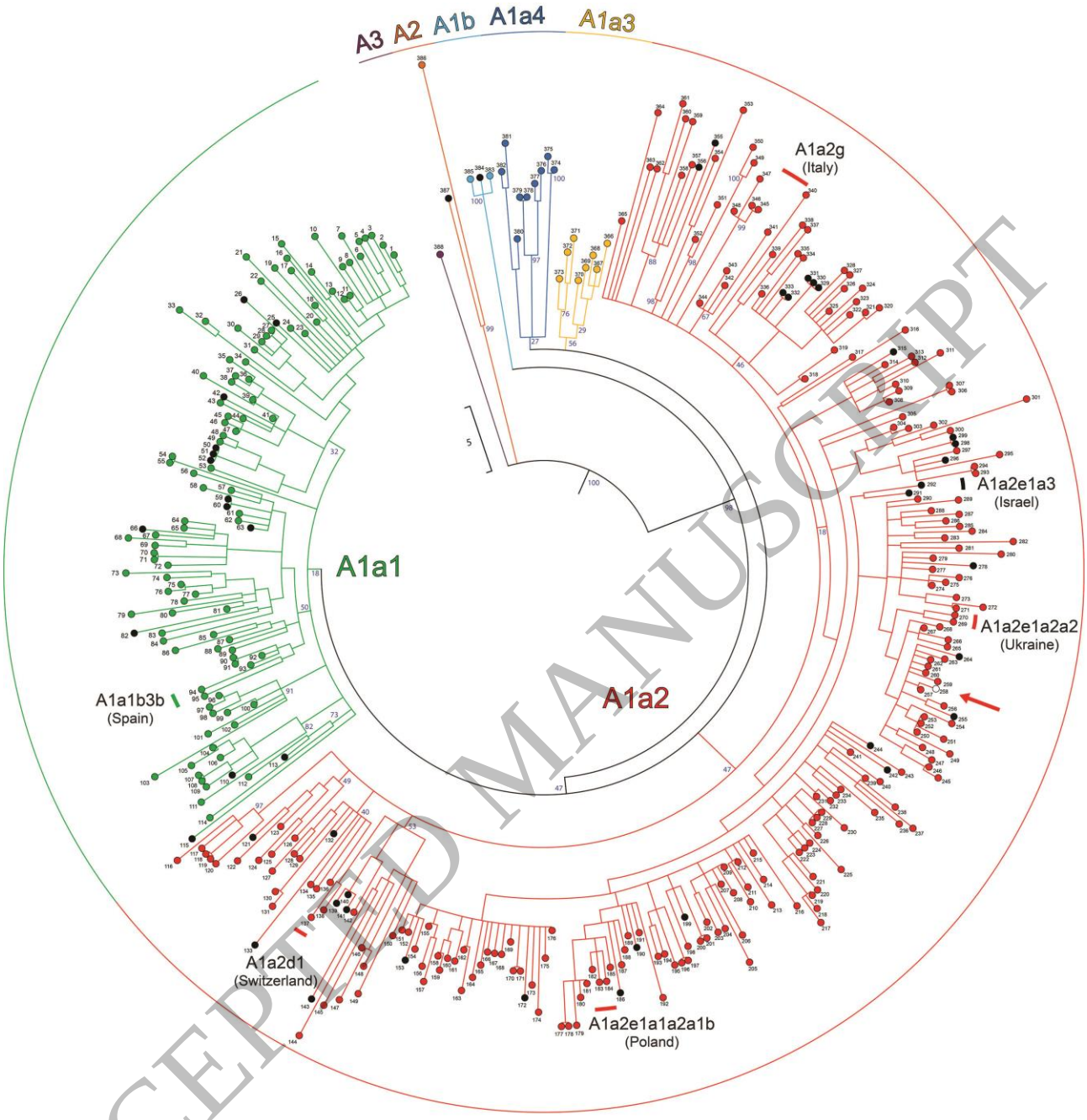


Figure 2
170x177 mm (.76 x DPI)

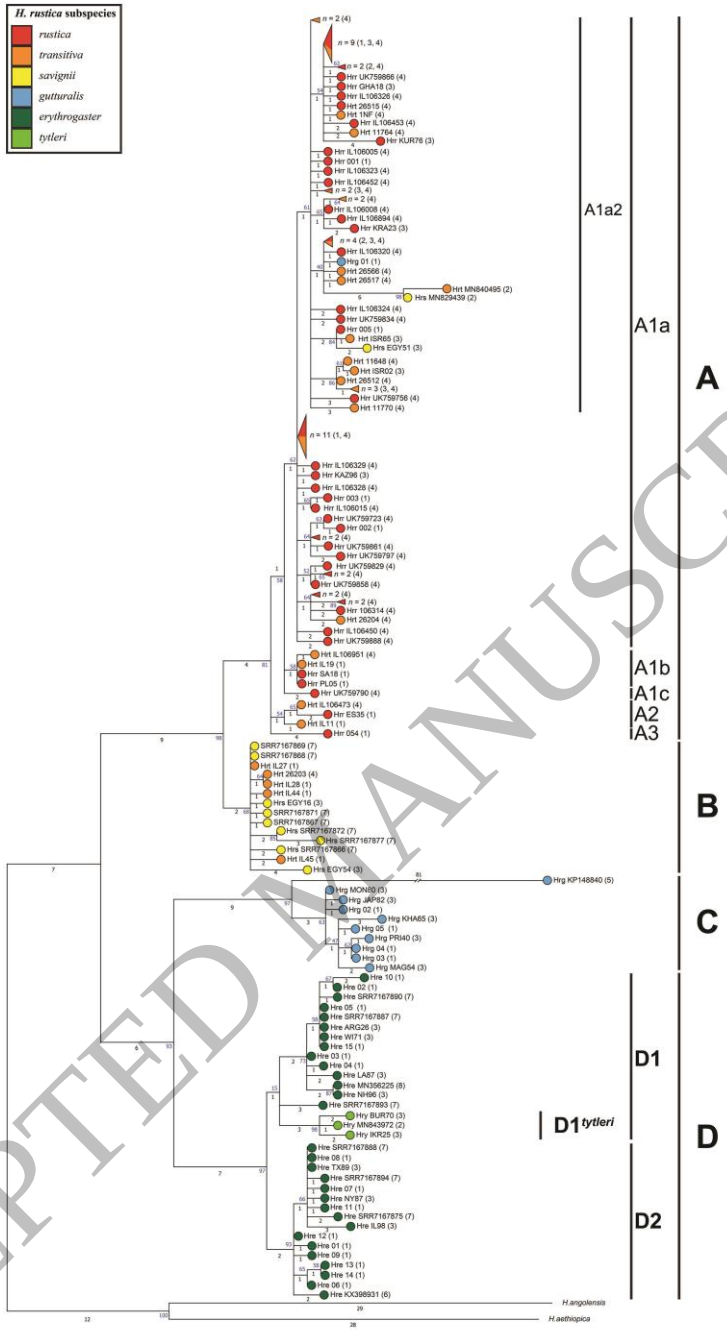


Figure 3
100x181 mm (.76 x DPI)

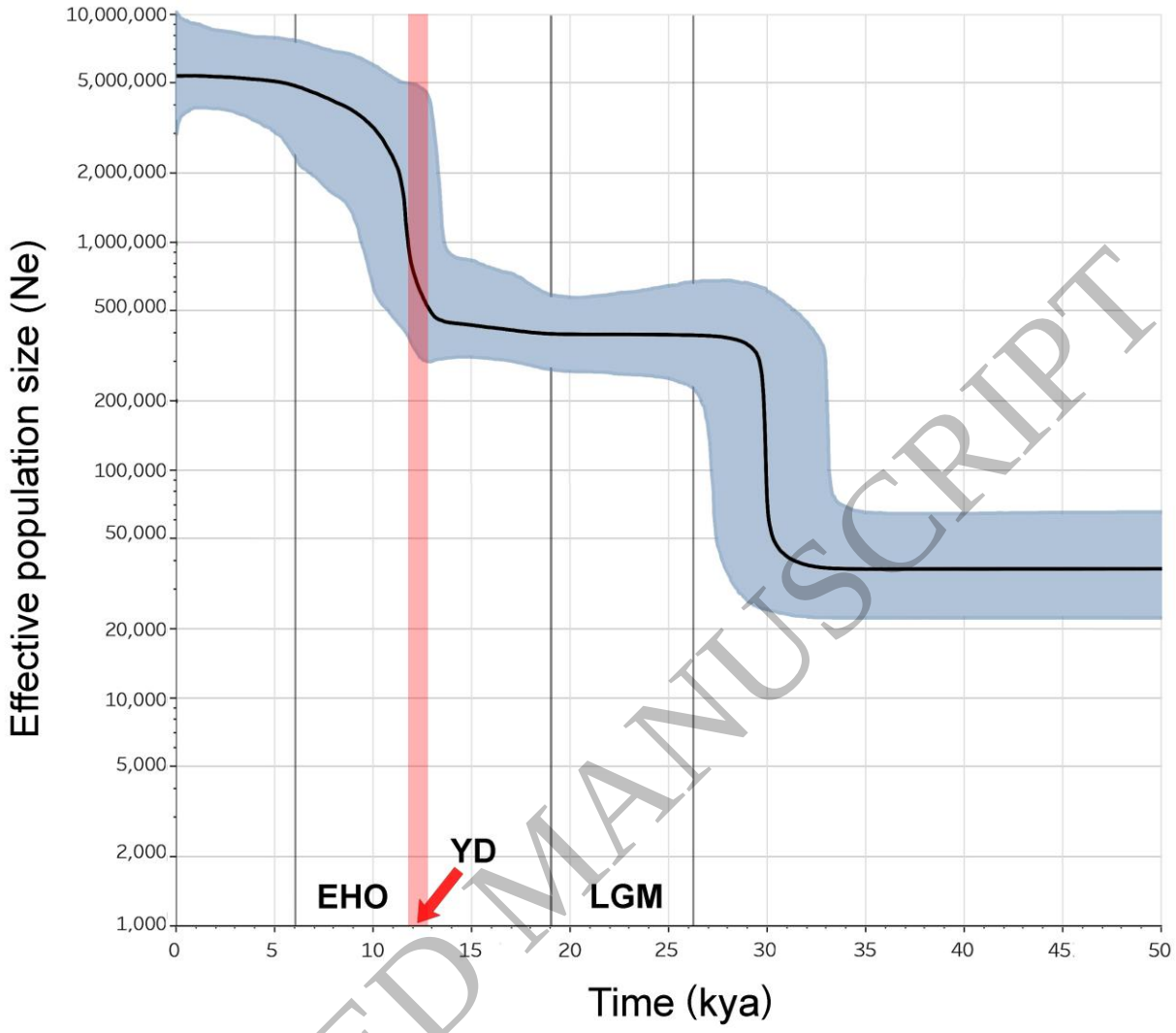


Figure 4
170x143 mm (.76 x DPI)

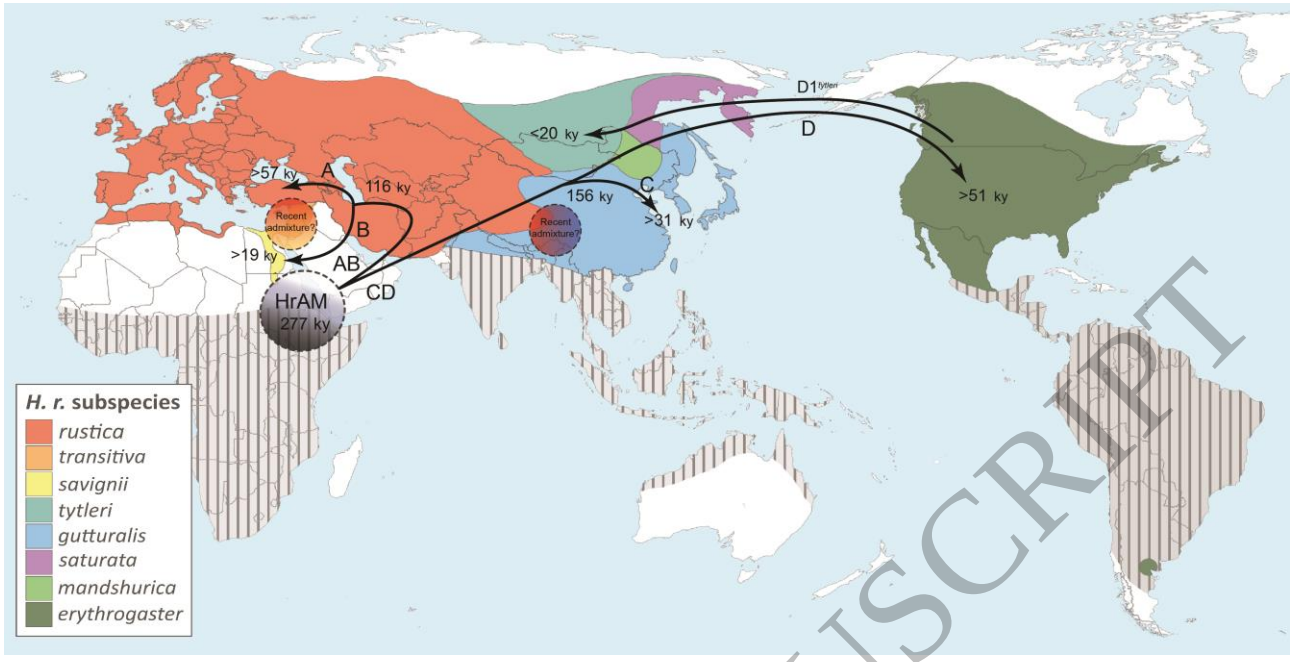


Figure 5
170x86 mm (.76 x DPI)

Chapter 4

Secomandi et al. (2021) **"The genome sequence of the European nightjar, *Caprimulgus europaeus* (Linnaeus, 1758)"**. *Wellcome Open Research*.

DATA NOTE

The genome sequence of the European nightjar, *Caprimulgus europaeus* (Linnaeus, 1758) [version 1; peer review: 2 approved]

Simona Secomandi[1], Fernando Spina[2], Giulio Formenti[3,4], Guido Roberto Gallo[1], Manuela Caprioli[5], Roberto Ambrosini[5], Sara Riello[6], Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]Department of Biosciences, University of Milan, Milan, Italy
[2]Institute for Environmental Protection and Research (ISPRA), Ozzano dell'Emilia, Italy
[3]Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA
[4]Howard Hughes Medical Institute, Chevy Chase, MD, USA
[5]Department of Environmental Sciences and Policy, University of Milan, Milan, Italy
[6]Riserva Naturale Statale "Isole di Ventotene e S. Stefano", Ventotene, Italy

v1 **First published:** 07 Dec 2021, **6**:332
https://doi.org/10.12688/wellcomeopenres.17451.1
Latest published: 07 Dec 2021, **6**:332
https://doi.org/10.12688/wellcomeopenres.17451.1

Abstract

We present a genome assembly from an individual female *Caprimulgus europaeus* (the European nightjar; Chordata; Aves; Caprimulgiformes; Caprimulgidae). The genome sequence is 1,178 megabases in span. The majority of the assembly (99.33%) is scaffolded into 37 chromosomal pseudomolecules, including the W and Z sex chromosomes.

Keywords

Caprimulgus europaeus, European nightjar, Eurasian nightjar, genome sequence, chromosomal

This article is included in the Tree of Life gateway.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 1 07 Dec 2021	✓ report	✓ report

1. **Anne-Lyse Ducrest**, University of Lausanne, Lausanne, Switzerland

2. **Joshua Peñalba**, Museum für Naturkunde, Berlin, Germany

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Secomandi S: Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Spina F**: Writing – Review & Editing; **Formenti G**: Conceptualization, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Gallo GR**: Resources, Writing – Review & Editing; **Caprioli M**: Writing – Review & Editing; **Ambrosini R**: Writing – Review & Editing; **Riello S**: Resources, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (206194) and the Darwin Tree of Life Discretionary Award (218328).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Secomandi S, Spina F, Formenti G *et al.* **The genome sequence of the European nightjar, *Caprimulgus europaeus* (Linnaeus, 1758) [version 1; peer review: 2 approved]** Wellcome Open Research 2021, **6**:332 https://doi.org/10.12688/wellcomeopenres.17451.1

First published: 07 Dec 2021, **6**:332 https://doi.org/10.12688/wellcomeopenres.17451.1

Species taxonomy

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Caprimulgimorphae; Caprimulgiformes; Caprimulgidae; Caprimulginae; Caprimulgus; *Caprimulgus europaeus* Linnaeus 1758 (NCBI: txid85660).

Background

The European nightjar (*Caprimulgus europaeus*; also known as the Eurasian nightjar and common goatsucker) is an insectivorous, crepuscular, ground-nesting bird distributed throughout the Western Palearctic (Hagemeijer & Blair, 1997). It breeds in semi-natural dry and open habitats with scattered trees (Cramp & Brooks, 1985). Little is known about the ecology of the European nightjar (Cramp & Brooks, 1985; Polakowski *et al.*, 2020), and in general that of the Caprimulgidae family. The family comprises peculiar species such as the only bird known to hibernate, the Common Poorwill (*Phalaenoptilus nuttallii*) (Carey, 2019; French, 2019; Woods *et al.*, 2019), and one of the few birds that uses echo-localization, the South American Oilbird (*Steatornis caripensis*) (Brinkløv *et al.*, 2013). The European nightjar has been found to be more resistant to pathogens than other bird species (Jiang *et al.*, 2021). Although categorized as 'least concern' by the IUCN (IUCN, 2016), the European nightjar has experienced a steady population decline in the past decades, and is of conservation concern in Europe (Eaton *et al.*, 2015; Evens *et al.*, 2017; Keller *et al.*, 2010). The availability of a high-quality, chromosome-level reference genome will help to deepen the knowledge on the biology and evolution of this species, boosting studies on the genomics of the peculiar family of Caprimulgidae. Moreover, as genomic resources gain preheminence in conservation efforts (Allendorf, 2017; Fuentes-Pardo & Ruzzante, 2017; Supple & Shapiro, 2018), we expect that the reference genome presented here will help aid planning conservation actions for the European nightjar.

Genome sequence report

The genome was sequenced from a blood sample taken from a single female *C. europaeus* collected from a bird ringing station in Ventotene, Italy (latitude 40.79404, longitude 13.42777). A total of 87-fold coverage in Pacific Biosciences single-molecule long reads and 62-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 144 missing/misjoins and removed 31 haplotypic duplications, reducing the assembly length by 0.15% and the scaffold number by 21.94%, and increasing the scaffold N50 by 26.46%.

The final assembly has a total length of 1,178 Mb in 121 sequence scaffolds with a scaffold N50 of 83 Mb (Table 1). Of the assembly sequence, 99.3% was assigned to 37 chromosomal-level scaffolds, representing 35 autosomes (numbered by sequence length) and the W and Z sex chromosomes (Figure 1–Figure 4; Table 2). The assembly has a BUSCO (Simão *et al.*, 2015) completeness of 97.4% (single 96.9%, duplicated 0.6%) using the aves_odb10 reference set. While not fully phased, the assembly deposited is of one pseudo-haplotype. Contigs corresponding to the alternate haplotype have also been deposited.

Table 1. Genome data for *Caprimulgus europaeus*, bCapEur3.1.

Project accession data	
Assembly identifier	bCapEur3.1
Species	*Caprimulgus europaeus*
Specimen	bCapEur3
NCBI taxonomy ID	NCBI:txid111811
BioProject	PRJEB44540
BioSample ID	SAMEA7524394
Isolate information	Female, blood
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6445211
10X Genomics Illumina	ERR6054683-ERR6054686
Hi-C Illumina	ERR6054687, ERR6054688
Genome assembly	
Assembly accession	GCA_907165065.1
Accession of alternate haplotype	GCA_907165095.1
Span (Mb)	1,178
Number of contigs	274
Contig N50 length (Mb)	31
Number of scaffolds	121
Scaffold N50 length (Mb)	83
Longest scaffold (Mb)	126
BUSCO* genome score	C:97.4%[S:96.9%, D:0.6%],F:0.5%,M:2.1%,n:8338

*BUSCO scores based on the aves_odb10 BUSCO set using v5.1.2. C= complete [S= single copy, D=duplicated], F=fragmented, M=missing, n=number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/bCapEur3.1/dataset/CAJRAV01/busco.

Methods

Sample acquisition

Sampling was performed during the routine activity of the scientific ringing station located in Ventotene island, Latina, Italy (latitude 40.7926°, longitude 13.4241°) during spring migration. Samples have been collected by ISPRA researchers within their institutional activities as from Italian national Law n. 157/92. Bird capture was performed in the evening according to standardized protocols using mist-nets (Saino *et al.*, 2010; Spina *et al.*, 1993). The sample was collected with a heparinized capillary tube after puncturing the ulnar

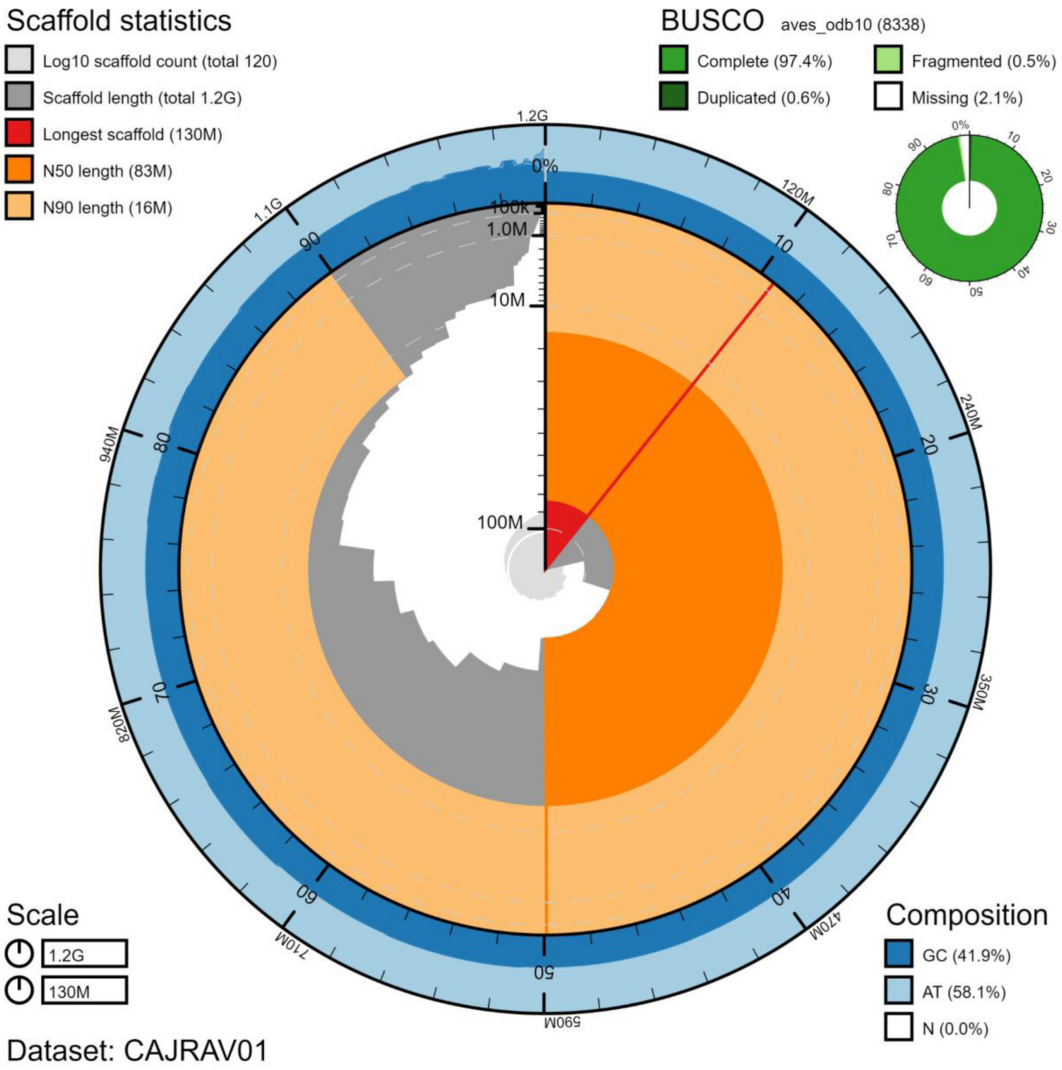


Figure 1. Genome assembly of *Caprimulgus europaeus*, bCapEur3.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 1,177,791,212 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (126,318,510 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (82,614,289 and 15,699,869 bp), respectively. The pale grey spiral shows the cumulative chromosome count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the aves_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/bCapEur3.1/dataset/CAJRAV01/snail.

vein with an intra-epidermal needle. The blood was immediately transferred into 99% ethanol, initially kept at room temperature and then frozen.

DNA extraction and sequencing

High molecular weight DNA was extracted from the blood sample at the Scientific Operations core of the Wellcome Sanger Institute using the Bionano Prep Blood DNA Isolation Kit according to the Bionano Prep Frozen Blood protocol. Pacific Biosciences CLR long read and 10X Genomics read cloud sequencing libraries were constructed according to the manufacturers' instructions. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II and Illumina HiSeq X instruments. Hi-C data were generated from the same blood sample using the Arima Hi-C+ kit and sequenced on HiSeq X.

Genome assembly

Assembly was carried out following the Vertebrate Genome Project pipeline v1.6 (Rhie *et al.*, 2020) with Falcon-unzip (Chin *et al.*, 2016), haplotypic duplication was identified and

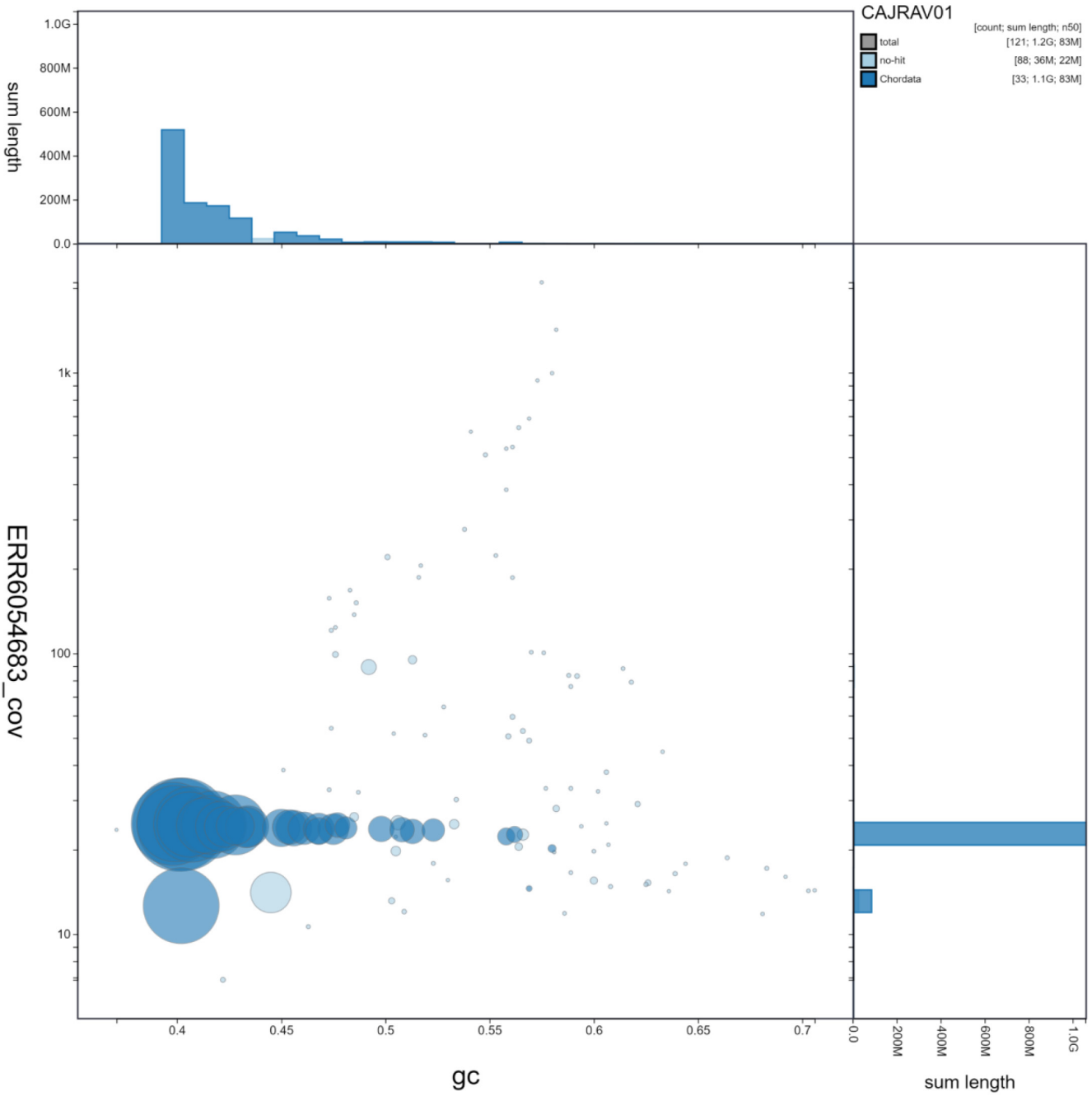


Figure 2. Genome assembly of *Caprimulgus europaeus*, bCapEur3.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/bCapEur3.1/dataset/CAJRAV01/blob.

removed with purge_dups (Guan *et al.*, 2020) and a first round of scaffolding carried out with 10X Genomics read clouds using scaff10x. Scaffolding with Hi-C data (Rao *et al.*, 2014) was carried out with SALSA2 (Ghurye *et al.*, 2019). The Hi-C scaffolded assembly was polished with arrow using the PacBio data, with merfin (Formenti *et al.*, 2021b) applied to avoid a drop in QV, then polished with the 10X Genomics Illumina data by aligning to the assembly with longranger align, calling variants with freebayes (Garrison & Marth, 2012) and applying homozygous non-reference edits using bcftools consensus. A complete mitochondrion was not found using mitoVGP (Formenti *et al.*, 2021a), likely due to the sample being sourced from blood tissue, so mitochondrial sequence NC_025773.1 (*Caprimulgus indicus*) was used during

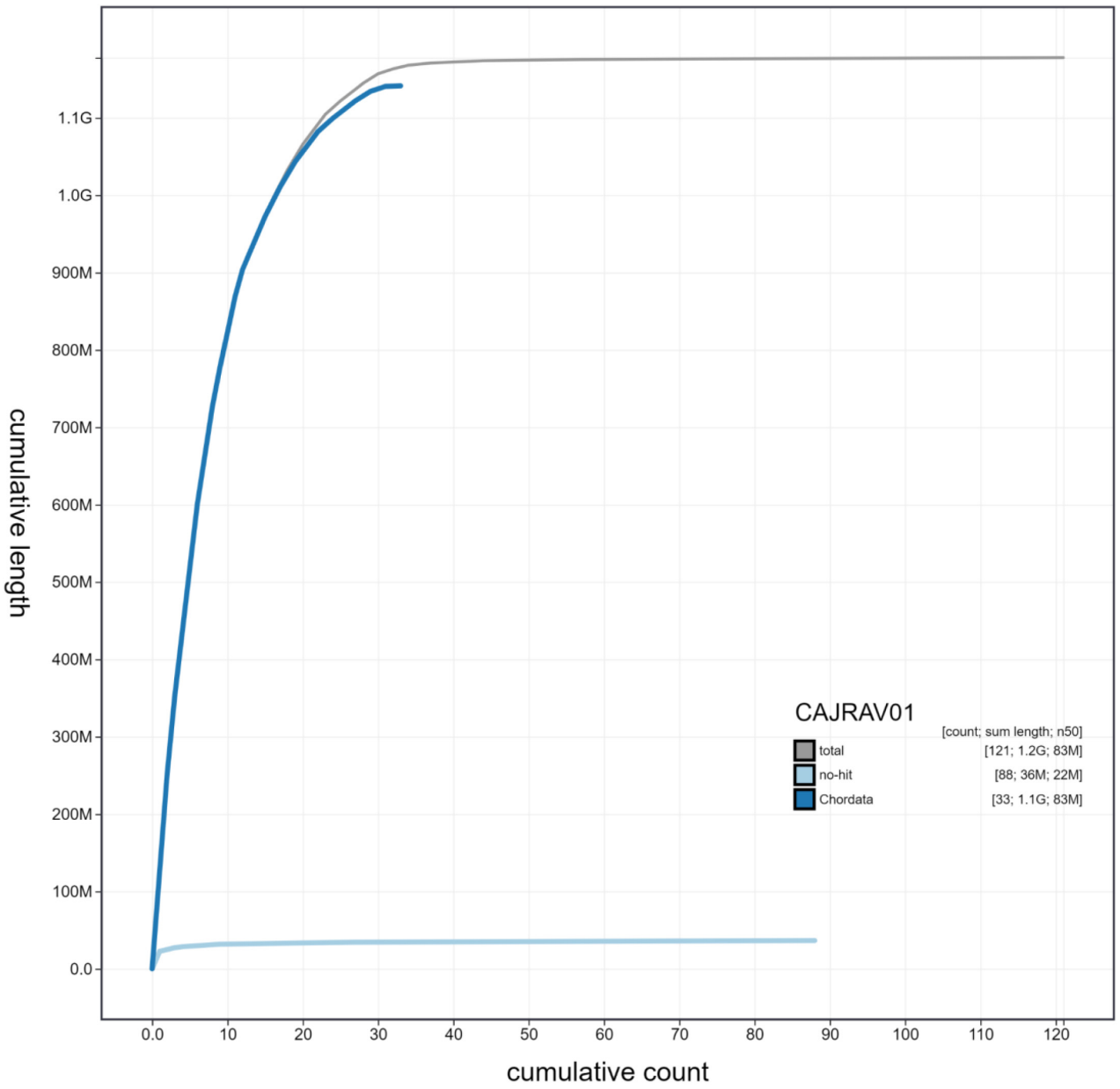


Figure 3. Genome assembly of *Caprimulgus europaeus*, bCapEur3.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/bCapEur3.1/dataset/CAJRAV01/cumulative.

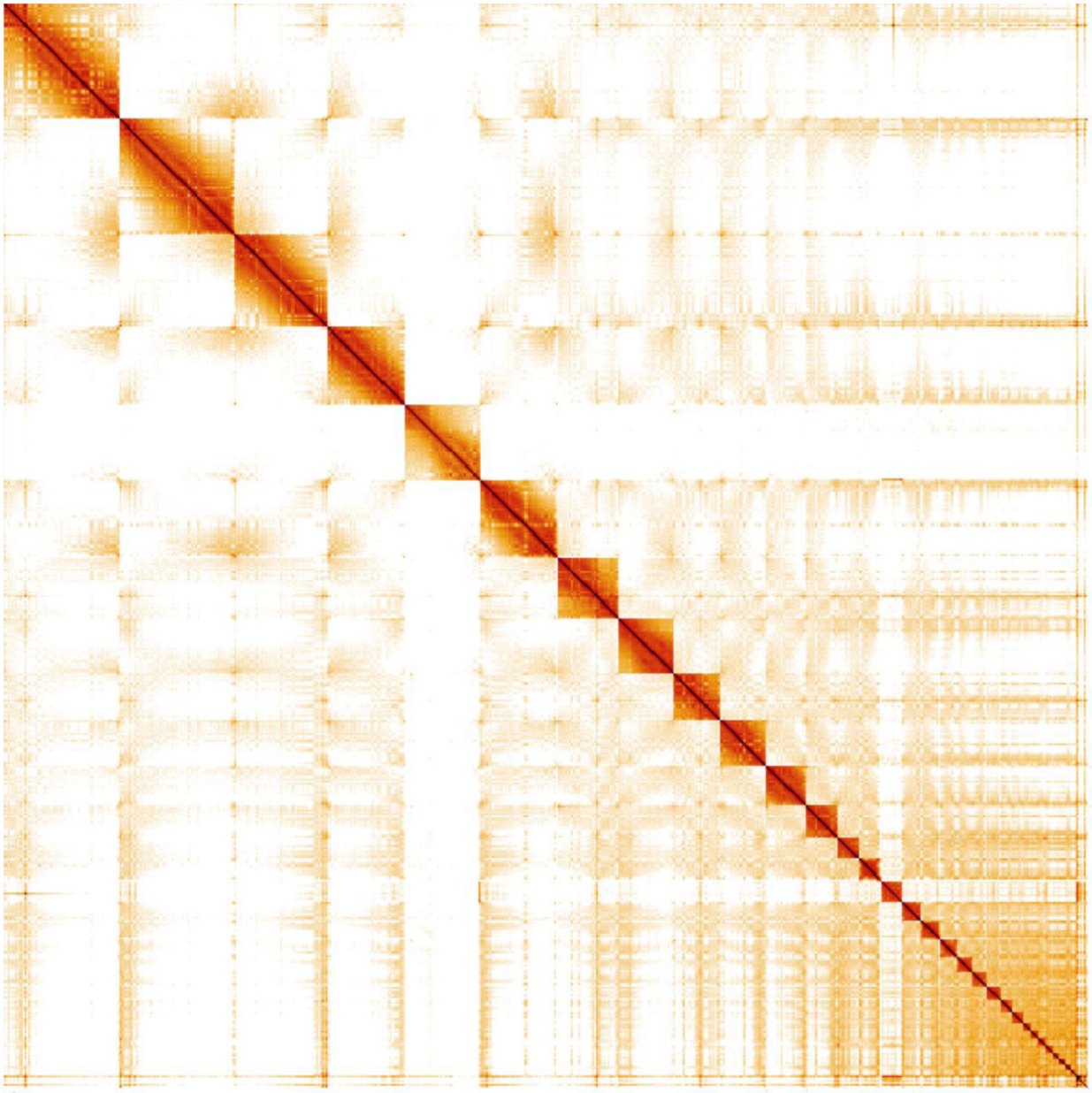


Figure 4. Genome assembly of *Caprimulgus europaeus*, bCapEur3.1: Hi-C contact map. Hi-C contact map of the bCapEur3 assembly, visualised in HiGlass. Chromosomes are shown in order of size from left to right and top to bottom.

polishing. The assembly was checked for contamination and corrected using the gEVAL system (Chow *et al.*, 2016) as described previously (Howe *et al.*, 2021). Manual curation (Howe *et al.*, 2021) was performed using gEVAL, HiGlass (Kerpedjiev *et al.*, 2018) and Pretext. The genome was analysed, and BUSCO scores generated, within the BlobToolKit environment (Challis *et al.*, 2020). Table 3 gives version numbers of the software tools used in this work.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Caprimulgus europaeus*, bCapEur3.1.

INSDC accession	Chromosome	Size (Mb)	GC%
OU015523.1	1	126.32	40.1
OU015524.1	2	125.37	40.3
OU015525.1	3	100.16	39.8
OU015526.1	4	83.32	39.9
OU015528.1	5	82.61	40.7
OU015529.1	6	65.35	41.7
OU015530.1	7	60.47	40.6
OU015531.1	8	50.91	42.8
OU015532.1	9	48.66	41.6
OU015533.1	10	43.00	41.3
OU015534.1	11	35.23	42.1
OU015535.1	12	23.52	43.4
OU015536.1	13	22.81	42.3
OU015538.1	14	22.35	43.3
OU015539.1	15	19.40	42.8
OU015540.1	16	18.74	45
OU015541.1	17	16.93	45.6
OU015542.1	18	15.70	45.4
OU015543.1	19	13.78	46.1
OU015544.1	20	12.52	46.8
OU015545.1	21	12.35	47.5
OU015546.1	22	9.16	46.8
OU015547.1	23	8.19	49.8
OU015548.1	24	7.57	47.7
OU015549.1	25	7.54	51.3
OU015550.1	26	7.50	50.8
OU015551.1	27	6.26	52.3
OU015552.1	28	6.04	48.1
OU015553.1	29	3.39	55.8
OU015554.1	30	2.94	56.1
OU015555.1	31	2.47	49.2
OU015556.1	32	2.22	50.6
OU015557.1	33	1.26	56.6
OU015558.1	34	0.56	51.3
OU015559.1	35	0.20	47.7
OU015537.1	W	22.49	44.5
OU015527.1	Z	82.63	40.2
-	Unplaced	7.86	54.9

Table 3. Software tools used.

Software tool	Version	Source
Falcon-unzip	1.8.0	Chin *et al.*, 2016
purge_dups	1.2.3	Guan *et al.*, 2020
SALSA2	2.2	Ghurye *et al.*, 2019
Arrow	GCpp-1.9.0	https://github.com/PacificBiosciences/GenomicConsensus
Merfin	1.7	Formenti *et al.*, 2021b
longranger align	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
freebayes	1.3.1-17-gaa2ace8	Garrison & Marth, 2012
gEVAL	N/A	Chow *et al.*, 2016
HiGlass	1.11.6	Kerpedjiev *et al.*, 2018
PretextView	0.1.x	https://github.com/wtsi-hpag/PretextView
BlobToolKit	2.6.2	Challis *et al.*, 2020

Data availability

European Nucleotide Archive: Caprimulgus europaeus (Eurasian nightjar). Accession number PRJEB44830; https://identifiers.org/ena.embl:PRJEB44830.

The genome sequence is released openly for reuse. The *C. europaeus* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project and the Vertebrate Genomes Project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1.

Author information

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783559.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893704.

Members of the Wellcome Sanger Institute Tree of Life collective are listed here: https://doi.org/10.5281/zenodo.4783586.

Members of the Sanger Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790456.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013542.

References

Allendorf FW: **Genetics and the Conservation of Natural Populations: Allozymes to Genomes.** *Mol Ecol.* 2017; **26**(2): 420–30.
PubMed Abstract | Publisher Full Text

Brinkløv S, Fenton MB, Ratcliffe JM: **Echolocation in Oilbirds and Swiftlets.** *Front Physiol.* 2013; **4**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

Carey C: **Life In The Cold: Ecological, Physiological, and Molecular Mechanisms**. CRC Press, 2019.
Publisher Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit - Interactive Quality Assessment of Genome Assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–74.
PubMed Abstract | Publisher Full Text | Free Full Text

Chin CS, Peluso P, Sedlazeck FJ, *et al.*: **Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing.** *Nat Methods.* 2016; **13**(12): 1050–54.
PubMed Abstract | Publisher Full Text | Free Full Text

Chow W, Brugger K, Caccamo M, *et al.*: **gEVAL — a Web-Based Browser for Evaluating Genome Assemblies.** *Bioinformatics.* 2016; **32**(16): 2508–10.
PubMed Abstract | Publisher Full Text | Free Full Text

Cramp S, Brooks DJ: **Vol. IV: Terns to Woodpeckers**. 1985.
Reference Source

Eaton M, Aebischer N, Brown A, *et al.*: **Birds of Conservation Concern 4: The Population Status of Birds in the UK, Channel Islands and Isle of Man.** *British Birds; an Illustrated Magazine Devoted to the Birds on the British List.* 2015; **108**(12): 708–46.
Reference Source

Evens R, Beenaerts N, Witters N, *et al.*: **Study on the Foraging Behaviour of the European Nightjar *Caprimulgus Europaeus* Reveals the Need for a Change in Conservation Strategy in Belgium.** *J Avian Biol.* 2017; **48**(9): 1238–45.
Publisher Full Text

Formenti G, Rhie A, Balacco J, *et al.*: **Complete Vertebrate Mitogenomes Reveal Widespread Repeats and Gene Duplications.** *Genome Biol.* 2021a; **22**(1): 120.
PubMed Abstract | Publisher Full Text | Free Full Text

Formenti G, Rhie A, Walenz BP, *et al.*: **Merfin: Improved Variant Filtering and Polishing via K-Mer Validation.** *bioRxiv.* 2021b.
Publisher Full Text

French AR: **Hibernation in Birds: Comparisons with Mammals**. In: *Life in the Cold*. CRC Press, 2019; 43–53.
Publisher Full Text

Fuentes-Pardo AP, Ruzzante DE: **Whole-Genome Sequencing Approaches for Conservation Biology: Advantages, Limitations and Practical Recommendations.** *Mol Ecol.* 2017; **26**(20): 5369–5406.
PubMed Abstract | Publisher Full Text

Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing**. arXiv: 1207.3907. 2012.
Reference Source

Ghurye J, Rhie A, Walenz BP, *et al.*: **Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly.** *PLoS Comput Biol.* 2019; **15**(8): e1007273.
PubMed Abstract | Publisher Full Text | Free Full Text

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–98.
PubMed Abstract | Publisher Full Text | Free Full Text

Hagemeijer WJM, Blair MJ: **The EBCC Atlas of European Breeding Birds.** Poyser, London, 1997; 479.
Reference Source

Howe K, Chow W, Collins J, *et al.*: **Significantly Improving the Quality of Genome Assemblies through Curation.** *GigaScience.* 2021; **10**(1): giaa153.
PubMed Abstract | Publisher Full Text | Free Full Text

IUCN: **Caprimulgus Europaeus: BirdLife International.** *IUCN Red List of Threatened Species*. IUCN, 2016.
Publisher Full Text

Jiang B, Zhenhua Z, Xu J, *et al.*: **Cloning and Structural Analysis of Complement Component 3d in Wild Birds Provides Insight into Its Functional Evolution.** *Dev Comp Immunol.* 2021; **117**: 103979.
PubMed Abstract | Publisher Full Text

Keller V, Gerber A, Schmid H, *et al.*: **Rote Liste Brutvögel. Gefährdete Arten Der Schweiz, Stand 2010. Umwelt-Vollzug Nr. 1019.** Bundesamt Für Umwelt, Bern, Und Schweizerische Vogelwarte, Sempach, 2010.
Reference Source

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
PubMed Abstract | Publisher Full Text | Free Full Text

Polakowski M, Broniszewska M, Kirczuk L, *et al.*: **Habitat Selection by the European Nightjar *Caprimulgus Europaeus* in North-Eastern Poland: Implications for Forest Management.** *Forests, Trees and Livelihoods.* 2020; **11**(3): 291.
Publisher Full Text

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.** *Cell.* 2014; **159**(7): 1665–80.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species.** *bioRxiv.* 2020; 2020.05.22.110833.
Publisher Full Text

Saino N, Rubolini D, Serra L, *et al.*: **Sex-Related Variation in Migration Phenology in Relation to Sexual Dimorphism: A Test of Competing Hypotheses for the Evolution of Protandry.** *J Evol Biol.* 2010; **23**(10): 2054–65.
PubMed Abstract | Publisher Full Text

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–12.
PubMed Abstract | Publisher Full Text

Spina F, Massi A, Montmaggiori A: **Spring Migration across Central Mediterranean: General Results from the" Progetto Piccole Isole.** 1993.
Reference Source

Supple MA, Shapiro B: **Conservation of Biodiversity in the Genomics Era.** *Genome Biol.* 2018; **19**(1): 131.
PubMed Abstract | Publisher Full Text | Free Full Text

Woods CP, Czenze ZJ, Brigham RM: **The avian "hibernation" enigma: thermoregulatory patterns and roost choice of the common poorwill.** *Oecologia.* 2019; **189**(1): 47–53.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 04 January 2022

https://doi.org/10.21956/wellcomeopenres.19297.r47480

✔ **Joshua Peñalba**
Center for Integrative Biodiversity Discovery, Museum für Naturkunde, Berlin, Germany

The authors describe the sequencing and assembly of the chromosome-scale reference genome for the European Nightjar. The methods follow that of the Vertebrate Genome Project pipeline. I just have some minor comments:

- How was the bird identified as female?
- About how much blood was used for the sequencing?
- How was the quality of the DNA checked?
- How many PacBio cells and Illumina lanes were used for each sequencing method?
- How did you know how many chromosomes should have been assembled?
- Can you provide more details on the assembly, which parameters were used and how was manual curation performed? If this is detailed in a different manuscript, please explicitly state which manuscript.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genomics, evolution, population genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 20 December 2021

https://doi.org/10.21956/wellcomeopenres.19297.r47481

Anne-Lyse Ducrest
Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

The authors described a nice almost complete genome with pseudo-chromosomes of the European nightjar using PacBio Sequel II, Illumina, and HiCi sequencing methods and thus present important data for further genetic analysis.

There are two points that could be improved:

- There are some redundancies between Figures 1, 2, and Table 1.
- The method how to get long HMV DNA is not well described since the Bionano protocol is for human blood and not for bird blood.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genomic, molecular biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Chapter 5

“A chromosome-level, haplotype resolved, reference genome for the lesser kestrel (*Falco naumanni*)”

Abstract

The availability of high-quality reference genome is pivotal to gain insights into the biology, ecology and evolution of a species. In this chapter I present a chromosome-level haplotype-phased reference genome for the lesser kestrel (*Falco naumanni*), a small migratory colonial kestrel distributed across southern Eurasia which is of European conservation concern. The genome was assembled using the VGP trio assembly pipeline v1.6. The new reference genome fulfilled the minimum VGP quality. 22 autosomes and the sex chromosomes (ZW) were identified, with the ~99% of the assembled sequences assigned to chromosomes. This reference will be included in a future paper with the main idea of determining how a migratory specialist like the lesser kestrel has coped with climatic fluctuations in the past and how it is expected to cope with them in the future under a scenario of climate change.

Introduction

The lesser kestrel (*Falco naumanni*) is a small (~120 g) migratory insectivorous falcon distributed across southern Eurasia[350]. In Europe, lesser kestrels breed in urban areas surrounded by farmlands, while Asian populations breed in natural habits[353]. Wintering areas are in sub-Saharan Africa[353]. The lesser kestrel is the target of many conservation actions and is currently classified as "least concern" by the IUCN Red List[360]. However, it is of European conservation concern and is listed in Annex 1 of European "Birds Directive" 2009/147/CE "SPEC 3" for BirdLife International 2017[361]. The lesser kestrel was the focus of many ecological[355,362–368], genetic[371,376–380] and conservation[369–375] studies. Genetic studies will be fostered by the generation of a complete and contiguous reference genome for the lesser kestrel. To this end, we generated a chromosome-level haplotype-resolved reference genome for the species using the VGP trio assembly pipeline v1.6[48]. We evaluated its quality, completeness and contiguity using the VGP standards. The reference genome is contributing to a study aimed at determining how a migratory specialist like the lesser kestrel has coped with climatic fluctuations in the past and how it is expected to cope with them in the future under a scenario of climate change (Bounas et al., forthcoming). To this end, whole-genome sequencing, mitogenomes, and ddRAD data analyses have been combined to infer past demographic trends, population structure and the phylogeographic history of the species. Past and recent gene flow across the species distribution were evaluated, and local adaptation using genotype-by-environment associations was assessed. From these analyses, a set of candidate genes for local adaptation has been selected and will be used to understand the relative importance of different

bioclimatic variables for local adaptation and estimate the genomic vulnerability (genetic offsets), which is an estimate of the risk of maladaptation, across their breeding distribution under different future scenarios of climate change. Evidence from demographic analyses, species distribution models (SDMs) from breeding and non-breeding distributions and risk of maladaptation based on adaptive variation will be combined to provide insights on the potential of the lesser kestrels to adapt to changing climatic conditions. In addition, the main differentiation among LK populations was found between the Asian and the European populations, which seem to be adapted to different ecological niches. While the range of the European clade is expected to increase due to climate change, the opposite trend is predicted for the Asian clade. The Asian clade has also a higher risk of maladaptation. Migratory distance is expected to increase, which could represent an additional challenge for the Asian populations. Finally, geographical isolation between the European and the Asian populations is expected to increase, which could promote reproductive isolation in the future and may lead to speciation. The VGP reference genome has been used for mapping the ddRAD sequencing reads, to infer the demographic history using MSMC2[381], to infer genome-wide heterozygosities and the annotation has been used to extract candidate genes from genotype-by-environment association analyses.

Results and discussion

A reference genome for the lesser kestrel. Using the VGP trio pipeline[48] we generate two assemblies for each haplotype of the lesser kestrel diploid genome starting from Illumina data generated from the parents and PacBio CLR reads, 10x Linked-Reads, Bionano optical maps and Hi-C reads generated from the child. Prior to the assembly process, a genome size of 1.34 Gbp, a repeat content of 298 Mbp and a heterozygosity of 0.89%, were predicted with Genomescope2.0[383] using *k*-mers from the 10x linked reads (**Figure 1a**, **Table 1**). At the end of the pipeline, the paternal haplotype assembly was chosen as the representative one based on its higher contiguity (scaffold N50 72.8Mbp vs 65.2Mbp, total scaffolds 491 vs 551). This haplotype was then subjected to manual curation along with the W sex chromosome (7 scaffolds, 2.96 Mbp) retrieved from the maternal haplotype. The combined assembly was run through purge_dups[384] to remove potential false duplications introduced by read misclassification during the trio-binning process. Following a manual review of the results, 158 scaffolds totalling 6.4 Mbp were removed, reducing the pre-curated assembly to 340 scaffolds and a total size of 1.22 Gbp. Finally, a round of manual curation performed at the end of the pipeline, introduced 98 rearrangements by breaking and joining scaffolds and resulted in the

removal of two additional false duplications. This process reduced the genome length by 4.99 Mbp, decreased the scaffold count by 15% to 290 and increased the scaffold N50 by 26% to 91.8 Mbp (**Table 2**). We will refer to the representative assembly as "bFalNau1" according to VGP guidelines for genome identifiers[48]. bFalNau1 and the maternal haplotype are publicly available on NCBI (GCF_017639655.2, GCA_017639645.1). With the VGP pipeline, we generated a 1.22 Gbp assembly divided into 290 scaffolds and 588 contigs, with a maximum scaffold length of 127 Mbp (**Table 2**). The genome size was lower than predicted (1.22 Gbp vs. 1.35 Gbp, **Figure 1a**, **Table 1,2**). However, the model fit from Genomescope2.0 was only 89% (**Table 1**), an unusually low value that could depend on the presence of some bias in 10x Linked-Reads that can result in somewhat inaccurate predictions. The assembly is highly contiguous, with a 87 Mbp scaffold NG50, a 12 Mbp contig NG50, and only 4 Mbp of gaps (298 gaps, N50 162 Kbp, **Table 2**). 22 autosomes and Z and W sex chromosomes were identified, assigning the 98.92% of the genomic sequence into chromosomes (**Figure 1b**). Only 32 scaffolds could not be placed in the corresponding chromosome and were submitted as "unlocalised". Our chromosomal reconstruction is in line with literature karyotypes for the *Falco* genus (2n = 40-52)[123]. bFalNau1 has a per-base assembly accuracy (QV) of 41.5 (less than 1 error every 10,000 bp), a *k*-mer completeness of 89.3% (**Table 3**), and a functional completeness of 96.8% based on the search of a set of *Vertebrata* orthologous genes performed with BUSCO[385] (**Figure 2**, **Table 4**). Of those genes, the 95.7% were single-copy and 1.1% duplicated. Moreover, only 1.2% of the found BUSCO genes were fragmented and 2% were missing (**Figure 2**, **Table 4**). The duplication content of the assembly was 0.07% (**Table 5**), lower than expected (1.42%, **Figure 1a**). Moreover, the *k*-mer count in the 10x Linked-Reads and their occurrence in the assembly did not show any significant content of duplicated sequences neither in the bFalNau1 assembly alone (Merqury[386] spectra-cn plots, **Figure 3a**), nor in the paternal (without W) and the maternal assembly combined (**Figure 3b**). The phasing completeness was confirmed by the blob-plot (**Figure 4a**) and phase blocks analyses (**Figure 4b**) generated with Merqury[386] on the paternal (without W chromosome) and the maternal assemblies. The trio assemblies had an NG50 phase block of 39 Mb (maternal) and 65.2 Mb (paternal) (**Table 6**). According to our results, the new bFalNau1 assembly fulfilled the minimum VGP quality "6.7.5.Q40.90" at the time of the assembly process (the notation is x.y.P.Q.C, where x = contig NG50 > 1 Mbp, y = scaffold NG50 > 10 Mbp, P = haplotype phased block NG50 > 100 kb, Q = QV base accuracy > 40, functional completeness > 90%, *k*-mer completeness > 90%, C = percentage of the assembly sequence assigned to chromosomes > 90%)[48].

Functional annotation. The NCBI Annotation Release 100 for the lesser kestrel identified 19,775 genes and pseudogenes. The alignment of the annotated proteins against a set of high-quality proteins (UniProtKB/Swiss-Prot), using the annotated proteins as the query and the high-quality proteins as the target, found that out of the 16,066 coding genes, 15,612 had a protein with an alignment covering 50% or more of the query and 11,066 had an alignment covering 95% or more of the query.

Figures

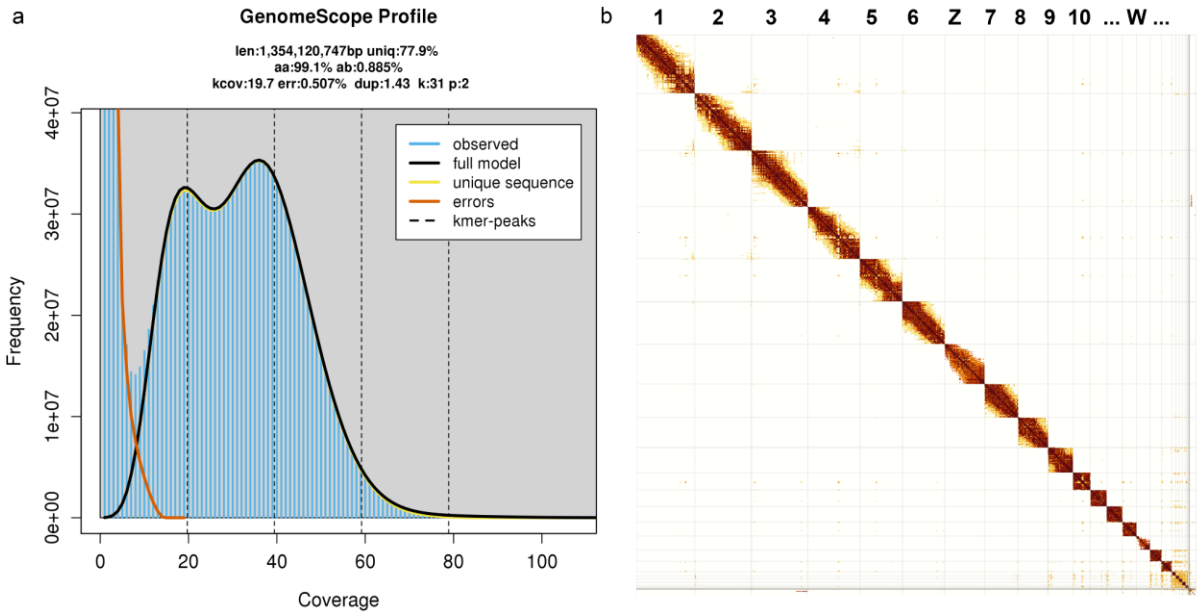


Figure 1. Genome assembly predictions and HiC heatmap. **a)** Linear plot from Genomescope2.0[383]. It reports the multiplicity of all the k-mers found in the 10x Linked-Reads (x-axis) versus their frequency (y-axis). This k-mers profile was used to measure and predict the genome size (len), the genome uniqueness (uniq), the homozygosity (aa) and heterozygosity (ab), the k-mer coverage (kcov) and the duplication content (dup). It also reports the error rate (err), the k-mer size used (k) and Genomescope version (p). **b)** Hi-C contact heatmap for the bFalNau1 assembly visualized with PretextView (https://github.com/wtsi-hpag/PretextView). The assembly is represented on both axes and between these shows the density of paired reads which fall into each interacting pair of regions in the plot. Colors discriminate the contact count. A scaffold is considered an end-to-end chromosome when all the reads that map to its sequence recreate a square, the diagonal is strong and there are no off-the-diagonal elements. Numbers at the top of the heatmap represent chromosome identifiers.

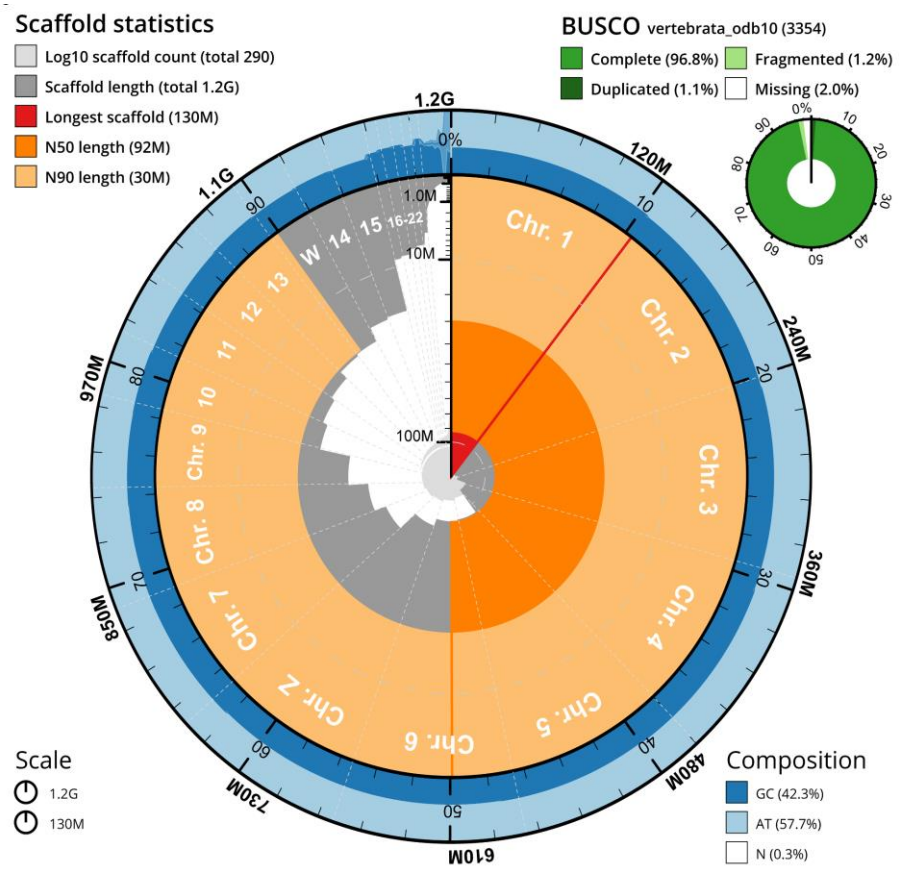


Figure 2. Snail plot summary of assembly statistics for bFalNau1 assembly. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 1.2 Gbp assembly. The distribution of scaffold lengths is shown in dark gray with the plot radius scaled to the longest scaffold present in the assembly (127 Mbp, in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (92 Mbp and 30 Mbp), 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 outside of the plot show the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes[385] in the vertebrata_odb10 set is shown in the top-right.

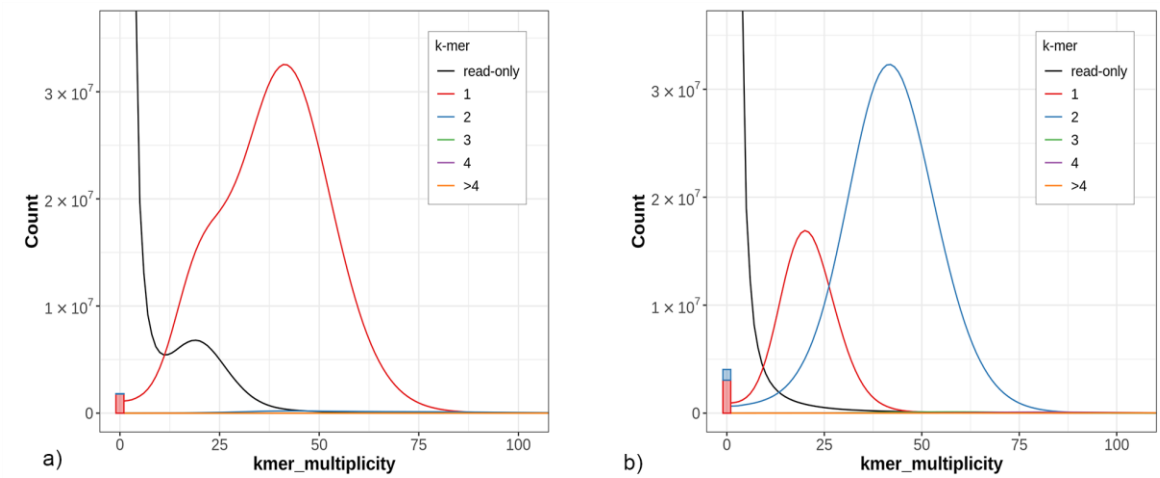


Figure 3. Merqury[386] spectra cn-plots. **a)** Merqury spectra-cn plots on the bFalNau1 assembly (paternal+W). It reports the multiplicity of the *k*-mers found in the 10x Linked-Reads (x-axis) and their frequency in the read set (y-axis). Coloured curves discriminate *k*-mers occurrence in the assembly. One main frequency peak (diploid peak) is visible at average sequencing depth (2-copy *k*-mers, homozygous or haplotype-specific duplications). A haploid peak at half average coverage (1-copy *k*-mers, homozygous) is also present and it includes the W sex chromosome specific *k*-mers. As expected, all *k*-mers were found only once in the assembly (red curve) since we are analyzing only the paternal haplotype. The black curve represents *k*-mers not found in the assembly. In particular, *k*-mers at low frequencies are almost always indicative of sequencing error (found only in the read set), while the 1-copy *k*-mers peaks are the ones belonging to the maternal haplotype. No extra-copy *k*-mers (artificial duplications) were detected. The bar at the origin of the graph represents *k*-mers found only in the assembly, which are considered assembly errors and are used by Merqury to compute the QV of the assembly. **b)** Merqury spectra-cn plot on the paternal haplotype (without the W chromosome) and the maternal haplotype combined. Two clear peaks are visible. The 1-copy *k*-mers at half average coverage were found once in the assemblies (only in one haplotype), while 2-copy *k*-mers were found twice (they are in both haplotypes or are haplotype-specific duplications) as expected, representing, therefore, complete haplotype-resolved assemblies. No significant extra-copy *k*-mers (artificial duplications) were detected.

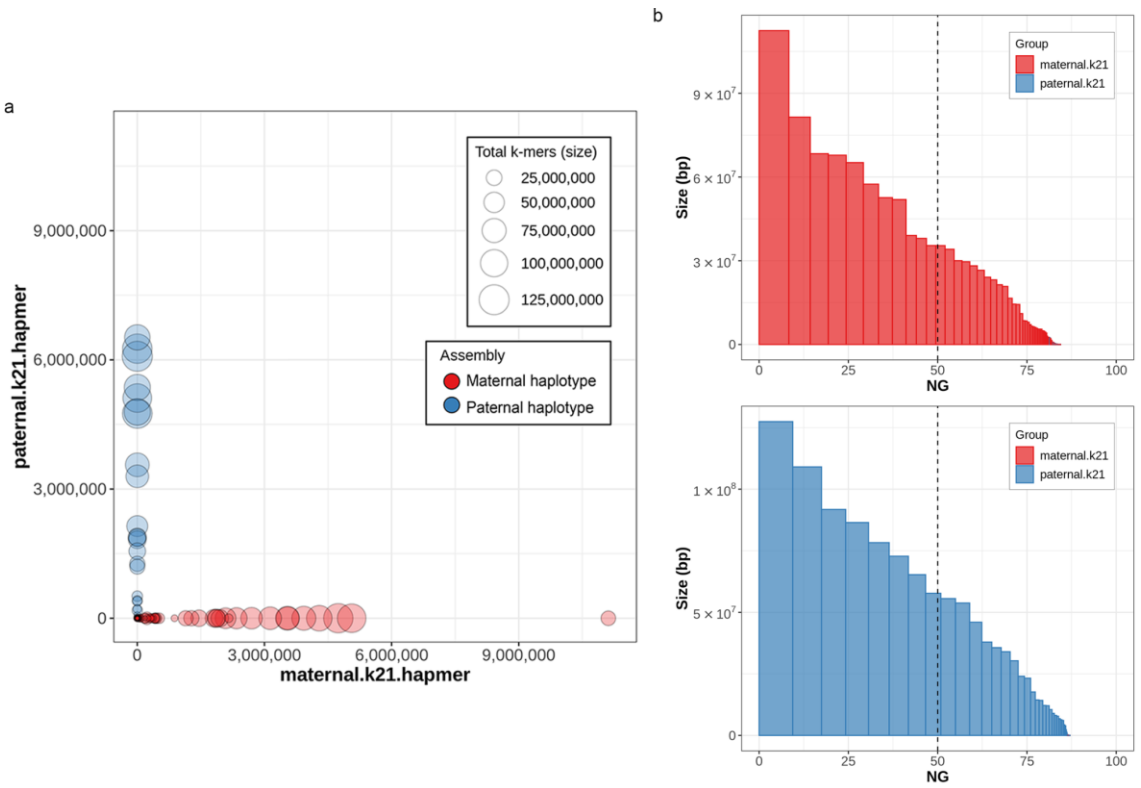


Fig 4. Merqury[386] hap-mer plots for evaluating haplotype phasing. **a)** Hap-mer blob plot of the paternal (W removed) and maternal assemblies to detect phasing consistency. Red blobs represent maternal scaffolds, while blue blobs are the paternal ones. Each blob is plotted according to the count of hap-mers found in the corresponding haplotype scaffold (maternal on the x-axis, paternal on the y-axis). Each scaffold was successfully separated by haplotype as expected. Indeed, both haplotype markers were observed in the corresponding assembly, with no contaminating markers from the alternate haplotype (haplotype-switches). Moreover, each blob (scaffold) is close to the corresponding assembly plot axis, representing therefore haplotype resolved assemblies. **b)** Phase block NG* plots of the haplotype resolved maternal assembly (red) and paternal assembly (with W chromosome removed, blue), sorted by size. The x-axis is the percentage of the genome size (*) covered by phase blocks of this size or larger (y-axis). Blocks from the wrong haplotype are almost entirely absent.

Tables

Property	min	max	mean
Homozygous (aa) (%)	99.1113	99.1186	99.11495
Heterozygous (ab) (%)	0.881429	0.888656	0.8850425
Genome Haploid Length (bp)	1,353,102,071	1,354,120,747	1,353,611,409
Genome Repeat Length (bp)	298,378,262	298,602,894	298,490,578
Genome Unique Length (bp)	1,054,723,809	1,055,517,853	1,055,120,831
Model Fit (%)	78.2897	99.7153	89.0025
Read Error Rate (%)	0.507461	0.507461	0.507461

Table 1. **Genomescope2.0**[383] **predictions on unassembled 10x linked reads for bFalNau1**. The table reports minimum and maximum estimations of genome homozygosity, heterozygosity, haploid length, repeated sequences length, unique sequences length, the model fit and the read error rate.

Statistics	Scaffolds	Contigs	Gaps
Total bp	1,215,702,009	1,211,624,722	4,077,287
Number	290	588	298
Max length (bp)	127,440,759	56,489,468	1,032,674
N50 (bp)	91,761,059	13,754,753	162,567
N90 (bp)	29,619,203	2,886,939	41,410
NG50 (bp)	86,597,978	12,381,245	-
NG90 (bp)	-	-	-

Table 2. **Assembly statistics calculated with asm_stats**. The table reports statistics related to assembled scaffolds, but also contigs and gaps.

Merqury stats	bFalNau1 (pat.+W)	pat.	mat.	mat. + pat.
QV	41.4534	41.4911	40.2164	40.8174
***k*-mer completeness (%)**	89.3	88.1597	83.5213	97.2819

Table 3. **QV and *k*-mer completeness computed with Merqury**[386] on the reference genome (paternal haplotype+W) and on the paternal and maternal haplotype alone and combined.

BUSCO genes	Number	%
Complete BUSCOs (C)	3246	96.8
Complete and single-copy BUSCOs (S)	3210	95.7
Complete and duplicated BUSCOs (D)	36	1.1
Fragmented BUSCOs (F)	40	1.2
Missing BUSCOs (M)	68	2.0
Total BUSCO groups searched	3354	100

Table 4. BUSCO gene prediction. The table reports BUSCO genes found in bFalNau1 starting from the vertebrate_odb10 database.

cutoff	1	2	3	4	>4	dup(>1)	all	dup%
62	1.07E+09	6958685	354418	63460	54668	7431231	1.08E+09	0.688172

Table 5. **False duplications content calculated with Merqury**[386]**.** The artificial duplication content of the assembly was 0.69%.

Assembly	Mat	Pat
Number of blocks	650	316
Total bases in blocks (Block sum)	1,142,583,379	1,178,113,705
Smallest block size	22	22
Avg. block size	1,757,821	3,728,208
Block N50 size	39,038,257	65,263,150
Longest block size	112,414,403	127,430,536
Number of markers from the other haplotype	15,190	12,369
Total number of markers in blocks	67,098,869	59,517,078
Switch error rate	0.02%	0.02%

Table 6. Phase blocks statistics with switch errors, allowing at most 100 switches within 20 kbp, calculated Merqury[386] for the parental trio assemblies using haplotype specific *k*-mers from parental data; the trio assemblies had NG50 phase blocks of 39 Mb (maternal) and 65.2 Mb (paternal).

Methods

DNA extraction. DNA extraction was performed with the Nanobind - Bionano Prep SP Frozen Human Blood DNA Isolation Protocol (Document Number: 30246, Document Revision: F, https://bionanogenomics.com/wp-content/uploads/2019/04/30246-Bionano-Prep-SP-Frozen-Human-Blood-DNA-Isolation-Protocol.pdf). Whole blood in ethanol was used for the DNA extraction. An aliquot was spun down and ethanol was removed. 5-10ul of nucleated blood was used. 1X PBS was added to blood for a total volume of 40ul. The sample was treated with Proteinase K and RNase A. A Nanobind Disk was used to bind gDNA. After three wash steps, the disk was transferred to an eppendorf tube and the gDNA was eluted with Buffer EB. The quality control was done through a pulsed field gel electrophoresis (PFGE) (Pippin Pulse, SAGE Science, Beverly, MA). According to the PFGE run, the length of the isolated DNA was > 200kb kbp.

Library preparation and sequencing. The assembly process involved four different sequencing technologies (Pacbio CLR long reads, 10x Genomics Linked-Reads, Bionano optical maps and Hi-C reads) generated from the nestling, plus Illumina short reads data for the mother and the father. For PacBio library preparation, 3.4µg of uHMW DNA was sheared using a 26G blunt end needle (Pacbio protocol PN 101-181-000 Version 05). A large-insert Pacbio library was prepared using the Pacific Biosciences Express Template Prep Kit v2.0 (#100-938-900) following the manufacturer protocol. The library was then size selected (>20kb) using the Sage Science BluePippin Size-Selection System. Then, the PacBio Library was sequenced on one PacBio 8M (#101-820-200) Smrtcell on the Sequel II instrument with the sequencing kit 2.0 (#101-820-200) using the Binding Kit 2.0 (#101-842-900) and a 15-hour movie. 10x Chromium libraries were regenerated from unfragmented HMW DNA from the agarose plugs on 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). The libraries were sequenced on an Illumina NovaSeq S4 150bp PE lane. For Bionano libraries, unfragmented uHMW DNA from the agarose plugs was labeled using a direct labeling enzyme (DLE1) following the DLS protocols (document Number 30206). The samples were then imaged on a Bionano Saphyr instrument. Chromatin interaction (Hi-C) libraries were generated on muscle with in-vivo cross-linking and sequenced on Illumina instruments. The preparations were performed by Arima Genomics (https://arimagenomics.com/) with the two-enzymes Arima-HiC kit (P/N: A510008). The proximally ligated DNA resulting from the kit was sheared, size-

selected (around ~200-600bp) using SPRI beads and enriched with streptavidin beads for biotin-labeled proximity-ligated DNA. The KAPA Hyper Prep kit (P/N: KK8504) was then used to generate Illumina-compatible libraries from those fragments. The libraries were amplified through PCR, purified with SPRI beads and checked for quality with qPCR and Bioanalyzer. Finally, the libraries were sequenced on Illumina HiSeq platforms following manufacturer's protocols. Illumina parental libraries were also prepared and sequenced on Illumina instruments.

Quality control. To detect species contamination and potential outlier sequencing runs, all sequencing data were evaluated with Mash[387] with 21 *k*-mers to generate sketches of size 10,000. No contamination was detected.

Genome size, heterozygosity and repeat content prediction from raw data. Using *k*-mer (genomic substrings of length *k*) based approaches, it is possible to estimate assembly characteristics from raw data before the assembly process. Genomescope2.0[383] was used to make predictions starting from the 31 bp *k*-mer profile generated with Meryl[386], a *k*-mer counting tool, from unassembled 10x Linked-Reads. The command used for generating the k-mer profile was: `./_submit_build_10x.sh 31 R1.fofn R2.fofn bFalNau1`. The script is part of the Merqury[386] GitHub repository (https://github.com/marbl/merqury/blob/master/_submit_build_10x.sh) and includes the trimming of the barcodes (first 23 bp of the first read pair), the building of a Meryl 31bp *k*-mer database. It generated a histogram of the *k*-mer counts that was used to compute estimations with Genomescope v2.0 (http://qb.cshl.edu/genomescope/genomescope2.0/)[383].

Assembly process. The lesser kestrel haplotype assemblies were generated with the VGP trio assembly pipeline v1.6[48] starting from genomic data generated from the nestling, the father and the mother. All the steps were run on the DNAnexus platform (https://www.dnanexus.com/). The nestling PacBio CLR long reads were binned to maternal and paternal haplotypes starting from the parental Illumina WGS short reads. TrioCanu v1.8[388] used the binned CLRs to assemble the two haplotypes separately as haplotigs, i.e. haplotype-specific contigs. The two sets of haplotigs were named mat_c1 and mat_c2. Both were processed independently following the same pipeline. First, the c1 haplotigs were polished with a step of Arrow (smrtanalysis 5.1.0.26412) and subjected to a purging step with purge_dups v1.0.0[384] to remove

retained alternate haplotigs (sequences belonging to the alternate haplotype, i.e. false duplications). The curated primary haplotigs (referred as p) were subjected to three scaffolding steps. Firstly, two rounds of scaffolding with 10x Linked-Reads were performed. The reads were aligned to the p intermediate and an adjacency matrix was computed with scaff10x v4.1.0 (https://github.com/wtsi-hpag/Scaff10X) from reads barcodes. The first round was run with `-matrix 2000 -reads 12 -link 10` parameters, and the second with `-matrix 2000 -reads 8 -link 10`. Contigs were joined into scaffolds separated by 100 bp gaps ("N"s). Secondly, the resulting scaffolds (called s1) were scaffolded again with Bionano optical maps, generated using the Bionano Pipeline, with Bionano Solve v3.2.1_04122018 using a DLE-1 one enzyme non-nicking approach. Gaps between contigs were sized according to the software estimate. Finally, the Hi-C reads previously binned to the parental haplotypes were aligned to the generated scaffolds (named s2) with the Arima Genomics mapping pipeline (https://github.com/ArimaGenomics/mapping_pipeline) to perform the last scaffolding step. Briefly, reads from each pair were aligned independently with BWA-MEM[389] with the `-B8` parameter and filtered for a minimum mapping quality of 10. Reads containing a restriction enzyme site (chimeric reads) were trimmed at the 3'. The single reads were then rejoined with their mate. The resulting alignments were used by Salsa2 HiC v2.2 for scaffolding[390] with parameters `-m yes -i 5 -p yes`, also indicating the restriction enzymes used for library generation (`-e GATC, GANT`). The final scaffolds (referred to as s3) were then subjected to a first step of polishing with Arrow (pacific Biosciences; smrtanalysis 5.1.0.26412) using binned CLR reads and renamed as t1. The alignment step was run with the command "`pbalign --minAccuracy=0.75 --minLength=50--minAnchorSize=12--maxDivergence=30--concordant--algorithm=blasr --algorithmOptions=--useQuality --maxHits=1 --hitPolicy=random --seed=1`", while the consensus polishing with "`variantCaller --skipUnrecognizedContigs haploid -x 5 -q 20 -X120 -v --algorithm=arrow`". The t1 intermediate was merged with the alternate haplotype scaffolds (also t1) and the mitogenome to perform the second polishing step with binned 10x Linked-Reads. Two rounds were performed, in which reads were aligned with Longranger 2.2.2[391], variants were called with FreeBayes v1.3.1[392] with default parameters and the consensus was generated with bcftools consensus[393] with `-i'QUAL>1 && (GT="AA" || GT="Aa")' -Hla`. The polished output was named as t2 after the first round and t3 after the second one. The t3 was then split into maternal and paternal haplotype (mat.asm, pat_asm).

Manual curation. Manual curation is pivotal to obtain high-quality chromosome-level reference assemblies. It involves contaminants and false duplications removal, the correction of assembly errors and the assignment of scaffolds into chromosomes. A decontamination pipeline, the genome evaluation browser gEVAL manual curation v. 2020-07-14 (geval.org.uk) and HiGlass Hi-C 2D maps were used as described in Howe et al, 2021[222].

Functional annotation. Newly-generated IsoSeq and RNAseq data, 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[48,394] was used for the annotation.

Assembly statistics. The assembly metrics were computed with the script asm_stats.sh, which is part of the VGP pipeline GitHub repository (https://github.com/VGP/vgp-assembly/blob/master/pipeline/stats/asm_stats.sh). The mean assembly size predicted with Genomescope2.0 (**Table 1**) was used in the command: `asm_stats.sh bFalNau1.fasta 1353611409 c`.

Repeat masking. The bFalNau1 assembly was masked with a combination of Windowmasker v1.0.1[395] and RepeatMasker 4.1.0[396]. Firstly, Windowmasker was used to soft mask the assembly. RepeatMasker was then ran on the assembly with NCBI/RMBLAST 2.10.0+ search engine using Dfam_3.1 (profile HMM library) and Rebase version 20170127[397], with "aves" as selected repeat library and the `-xsmall` parameter to perform soft-masking (repeats represented as lowercases). The masking coordinates from RepeatMasker were used on the WindowMasker masked assembly using `bedtools maskfasta -soft` to merge the two different sets of repeats.

Blobtoolkit. Blobtoolkit[398] was used to assess the quality of the assembly in an interactive way. The database for bFalNau1 was generated with blobtools v2.3.3: `blobtools create --fasta bFalNau1.fasta --taxid 8782 --taxdump path/to/taxdump db_name` and BUSCO results were added with `blobtools add --busco output.tsv db_name`.

Hi-C contact heatmaps. The three-dimensional conformation of chromosomes can be visualized with interaction heat maps obtained aligning the HiC the read set against the assembly. This analysis was performed on the DNAnexus platform (https://platform.dnanexus.com/). Briefly, the assembly was indexed with bwa fasta indexer 2.0.2 and the reads were aligned with bwa mem. PretextMap (https://github.com/wtsi-hpag/PretextMap) was used to generate a contact map that was visualized with PretextView (https://github.com/wtsi-hpag/PretextView).

Functional completeness assessment. To compute the functional completeness of the assembly, BUSCO 4.1.4[385,399,400] was used with a customized config file. BUSCO searches for a set of highly conserved orthologous genes in a genome. Missing, duplicated or fragmented BUSCO genes may reflect assembly errors or sequence incompleteness. The vertebrate_odb10 database (OrthoDB v10) was used to search a set of orthologous genes that are found in single copy across vertebrates, using "chicken" as the training species for Augustus 3.3.3, which is one of the integrated softwares, together with Hmmer 3.1b2 and BLAST 2.10.1+.

QV, k-mer completeness, duplication content. Merqury[386] is an evaluation tool that compares a set of *k*-mers derived from unassembled reads to a genome assembly to compute consensus quality (QV), *k*-mer completeness, duplication content and graphical visualization of the copy number spectrum and *k*-mer coverage across the assembly. Merqury can also report haplotype completeness when the parental genomes are available. The *k*-mer count was performed with Meryl[386], a *k*-mer counting tool extended to support Merqury operations from child's 10x linked reads. The command was the following:
`./_submit_build_10x_mod.sh 21 R1.fofn R2.fofn child`
(https://github.com/marbl/merqury/blob/master/_submit_build_10x.sh), where 21 is the selected *k*-mer size (21 bp), R1.fofn and R2.fofn links to the forward and reverse 10x Linked-Reads, respectively, and "child" is the output name. Using the obtained 21bp *k*-mers database, Merqury counted the occurrence of each *k*-mer in the assembly. It was run on both the reference assembly bFalNau1 (paternal+W) and the combination of paternal (without W) and maternal assemblies. To obtain the paternal assembly, the W chromosome was removed from the reference. The script used to perform the analysis was https://github.com/marbl/merqury/blob/master/_submit_merqury.sh and was ran with the following commands:

`_submit_merqury.sh k21.meryl reference.fasta`, for bFalNau1, and `_submit_merqury.sh k31.meryl paternal.fasta maternal.fasta`, for the trio assemblies. Spectra-cn plots were then generated in both cases. They track the *k*-mer multiplicity of each *k*-mer found in the unassembled reads in the assembly. A typical *k*-mer spectrum is composed of two peaks, the first representing *k*-mers found in 1 copy in the assembly (heterozygous), and the second *k*-mers found twice (homozygous or haplotype-specific duplications). The 2-copy *k*-mers are expected to appear at average depth of the sequencing coverage, while the 1-copy *k*-mers at half sequencing coverage. The duplication content was calculated with the script https://github.com/marbl/merqury/blob/master/eval/false_duplications.sh starting from the *k*-mer histogram generated with Merqury. Merqury also calculated the QV from the frequency of consensus errors in the assembly, and *k*-mer completeness, which is the fraction of reliable read *k*-mers in the assemblies (i.e. *k*-mers that are truly in the assembly and not caused by sequencing errors).

Phasing completeness assessment. Merqury[386] was also used for assessing the phasing completeness of the assemblies. Firstly, parents *k*-mer databases were obtained with Meryl from the maternal and paternal Illumina reads with the https://github.com/marbl/merqury/blob/master/_submit_build.sh script: `./_submit_build_mod.sh 21 input.fofn paternal` and `./_submit_build_mod.sh 21 input.fofn maternal`. The child database was already generated ("QV, k-mer completeness, duplication content" section in **Methods**). The haplotype specific markers (hap-mers) were obtained from the parental assemblies. Briefly, *k*-mers found only in one parent genome are collected and low-frequency *k*-mers are filtered out. The inherited hap-mers were also obtained from the intersection of the child's *k*-mers with the parental hap-mers set. The script https://github.com/marbl/merqury/blob/master/trio/hapmers.sh was used for this purpouses with the following command: `./_submit_hapmers.sh .maternal.k21.meryl paternal.k21.meryl child.k21.meryl`. Merqury analysis was then performed with https://github.com/marbl/merqury/blob/master/_submit_merqury.sh:

`./_submit_merqury.sh child.k21.meryl maternal.k21.inherited.meryl paternal.k21.inherited.meryl paternal.fasta maternal.fasta`.

Phasing-completeness was evaluated with a hap-mer blob plot generated from the count of the hap-mers found in each scaffold, and with phase block plots and stats, all generated by the Merqury script. Phased blocks were determined from hap-mers, where a block is composed by a set of maskers derived from the same haplotype. To account for minor base-level errors in the assembly, short-range switches were allowed to occur within a block (at most 100 switches within 20 kbp).

Chapter 6

Formenti et al. (2022) **“The era of reference genomes in conservation genomics”**. *Trends in Ecology and Evolution.*

Trends in
Ecology & Evolution

CellPress

Forum

The era of reference genomes in conservation genomics

Giulio Formenti,[1,29]
Kathrin Theissinger,[2,3,4,29]
Carlos Fernandes,[5,6,29]
Iliana Bista,[7,8]
Aureliano Bombarely,[9]
Christoph Bleidorn,[10]
Claudio Ciofi,[11]
Angelica Crottini,[12]
José A. Godoy,[13]
Jacob Höglund,[14]
Joanna Malukiewicz,[15]
Alice Mouton,[16]
Rebekah A. Oomen,[17,18]
Sadye Paez,[1] Per J. Palsbøll,[19,20]
Christophe Pampoulie,[21]
María J. Ruiz-López,[13]
Hannes Svardal,[22]
Constantina Theofanopoulou,[1]
Jan de Vries,[23]
Ann-Marie Waldvogel,[24]
Guojie Zhang,[25,26]
Camila J. Mazzoni,[27]
Erich D. Jarvis,[1]
Miklós Bálint,[2,4,28,*]
European Reference Genome Atlas (ERGA) Consortium[30,31]

Progress in genome sequencing now enables the large-scale generation of reference genomes. Various international initiatives aim to generate reference genomes representing global biodiversity. These genomes provide unique insights into genomic diversity and architecture, thereby enabling comprehensive analyses of population and functional genomics, and are expected to revolutionize conservation genomics.

Conservation, genomics, and reference genomes

In 2020 both the United Nations Biodiversity Summit and the European Environment Agency emphasized the accelerating global loss of biodiversity (https://www.un.org/pga/75/united-nations-summit-on-biodiversity/; https://www.eea.europa.eu/highlights/latest-evaluation-shows-europes-nature). We are in the sixth mass extinction. Although the primary route to preserving biodiversity comprises protection of species and restoration of habitats and ecosystems, genomics provides a rapidly expanding array of novel tools to characterize biodiversity and assist such conservation efforts. The need for immediate actions that help to reverse the current biodiversity decline has prompted national and international initiatives aimed at expanding the genomic reference resources available for biodiversity research and conservation across the tree of life (Box 1). Many of these efforts collectively contribute to the Earth BioGenome Project (EBP) that aims to catalog and characterize the genomes of all of Earth's eukaryotic biodiversity. A large and inclusive community of scientists has recently gathered as the European hub of the EBP to promote the generation of a European Reference Genome Atlas (ERGA; www.erga-biodiversity.eu). This initiative is building a pan-European open access infrastructure to streamline ethical and legally compliant sample and metadata collection [1], sequencing and **assembly** (see Glossary) [2], annotation [3], and release in public archives of high-quality genomic information, thus creating reference genomes for a wide variety of eukaryotic species (Box 1).

Reference genomes, by which we mean highly contiguous, accurate, and annotated genome assemblies, greatly enhance genomic studies, both experimentally and analytically [2,4]. A reference genome is a point representation of the structure and organization of the genome of a species. Similarly to type specimens in taxonomy, reference genomes serve as the standard for subsequent genomic studies [5]. To cost-efficiently unravel the genomic diversity of species, multiple conspecific individuals can be resequenced and aligned to available reference genomes instead of being assembled *de novo*. Thus, reference genomes provide a comprehensive and fundamental framework onto which genomic variation can be mapped to characterize and ultimately aid in preserving genetic diversity [4]. To this end, special attention should be paid to the origin of the individuals used as the reference because, if these are excessively divergent from the populations under study, this could compromise subsequent

Glossary

Assembly: a chromosome-level contiguous sequence of all chromosomes, often aided by genetic maps or other information.
Evolutionary distinct and globally endangered (EDGE) species: species of high conservation priority.
Genetic rescue: a mitigation strategy for restoring intraspecific genetic diversity and reducing extinction risks in small, isolated, or inbred populations through induced gene flow.
Heterozygote advantage: when a heterozygous genotype has a higher relative fitness compared to a homozygous dominant or homozygous recessive genotype.
Hybridization: interbreeding of individuals from genetically distinct lineages.
Inbreeding depression: reduced fitness in offspring as a result of inbreeding – mating between closely related individuals.
Introgression: gene flow between hybridizing populations or species by backcrossing hybrids with one or both parental populations.
Metagenomics and metatranscriptomics: sequencing of DNA or RNA-derived cDNA extracted from environmental and bulk samples.
Outbreeding depression: reduced fitness in offspring from mating between genetically divergent individuals.
Pangenome: the entire set of DNA sequences (or genes) of a species represented by the core genome and the accessory genome.
Phylogenomics: the inference of the phylogenetic relationships among different lineages of organisms from genome-wide data.
Reference genome: a contiguous and accurate genome assembly representative of a species in which the coordinates of genes and other important features are annotated. Current definitions of reference genome quality are given in [2] and https://www.earthbiogenome.org/assembly-standards.

Box 1. Sequencing the tree of life

International initiatives aimed at generating genomic resources, and particularly reference genomes, have flourished in recent years. Some focus on specific taxa, such as the Vertebrate Genomes Project, Bird Genome 10K Project, Bat1K Project, Global Invertebrate Genomics Alliance, 10 000 Plant Genomes Project, and 1000 Fungal Genomes project. Others focus on geographic regions, such as the California Conservation Genomics Project, Darwin Tree of Life for Britain and Ireland, Catalan Initiative for the Earth BioGenome Project in the Catalan territories, Endemixit in Italy, Norwegian Earth Biogenome Project, and SciLifeLab in Sweden, on applications such as the LOEWE Translational Biodiversity Genomics in Germany, or on ecological systems such as the Aquatic Symbiosis Genomics project. Collectively part of the Earth BioGenome Project (EBP), in Europe these initiatives are organized under the umbrella of the European Reference Genome Atlas (ERGA).

A genome atlas of European biodiversity

ERGA is a pan-European scientific response to the current threats to biodiversity. Approximately one fifth of the ~200 000 eukaryotic species present in Europe can be inferred to be at risk of extinction according to the International Union for Conservation of Nature (IUCN) Red List classification (this estimate only considers the assessed species; https://www.iucn.org/regions/europe/our-work/biodiversity-conservation/european-red-list-threatened-species).

ERGA aims to generate reference genomes of European eukaryotic species across the tree of life, including threatened, endemic, and keystone species, as well as pests and species important to agriculture, fisheries, and ecosystem function and stability. ERGA builds upon current genomic consortia in EU member states, EU Associated Countries, representatives of other countries within the European bioregion, and international collaborators. These reference genomes will address fundamental and applied questions in conservation, biology, and health. ERGA seeks to alert the EU about the potential of conservation genomics, and particularly the role of reference genomes, in biodiversity assessment, conservation strategies, and restoration efforts.

analyses. To overcome this issue, multiple conspecific genomes [6] can now be summarized in the **pangenome** of a species [7].

Until recently reference genomes have only been available for a handful of model organisms. Thanks to the consolidated and standardized efforts of international genome initiatives, the situation is rapidly changing. Recent technological advances provide a general strategy for generating chromosome-scale reference genomes for all organisms across the tree of life [2]. These advances rely on a combination of single-molecule long-read sequencing [either PacBio Single Molecule Real-Time (SMRT) sequencing or Oxford Nanopore Technologies (ONT) sequencing] and/or linked reads [(e.g., transposase enzyme linked long-read sequencing (TELL-seq) or single-tube long fragment read (stLFR) sequencing] for contig assembly, optical mapping, and/or proximity ligation followed by high-throughput sequencing (Hi-C) for scaffolding [2].

Decreasing costs, improved scalability, and increasing quality of sequencing technologies, combined with better algorithms and advances in computational power [2], facilitate the establishment of reference genomes across the full spectrum of biodiversity. Importantly, reference genomes are fundamental for a comprehensive and accurate characterization of genomic information, for instance of structural features that cannot be inferred from fragmented genomes or reduced-representation sequencing approaches (Figure 1). Therefore, reference genomes coupled with resequencing data should become a standard in conservation genomics, facilitated by constantly evolving analytical methods.

Key contributions of reference genomes in conservation genomics

The full spectrum of genomic diversity

Reference genomes provide a view of the architecture of the genome, comprising both genic and intergenic regions. These include repetitive regions, some of which are challenging to assemble, such as segmental duplications, centromeres and telomeres, satellites, and mobile elements. Population genomics guided by reference genomes aids the identification of classical genetic variants, such as SNPs and copy number variants (CNVs), as well as structural variants that are particularly difficult to detect in fragmented and incomplete reference genomes alone, but are potentially important in adaptation to environmental change [8].

Inbreeding and deleterious mutations

Assessments of inbreeding have long informed conservation and breeding programs, guiding genetic crosses and translocations of individuals. Although often estimated from a few loci, understanding the genetic architecture and accurately quantifying inbreeding and **inbreeding depression** require a genome-wide perspective, encompassing for example the number of genes involved, the presence of alleles with large effects, the role of deleterious recessive alleles, and **heterozygote advantage** [9]. Although several questions remain, multiple studies have showcased the power of population genomics guided by reference genomes to identify runs of homozygosity as a means to estimate inbreeding, as well as to reveal the dynamics and fate of deleterious variation in threatened species (e.g., [10]).

Outbreeding and introgression

Mating between individuals from genetically distinct lineages may lead to **outbreeding depression** due to chromosomal or genic incompatibilities, epistatic interactions, disruption of interactions between co-adapted genes, or the introduction of maladaptive variants into local populations. Population genomics guided by reference genomes greatly aids the disentanglement of these phenomena [11]. **Hybridization** is a common evolutionary process that, through **introgression**, can promote the spread of adaptive variation and speciation. Anthropogenic hybridization and introgression, however, can be major threats to biodiversity and evolutionary heritage. Reference genomes facilitate the characterization of introgression patterns and dynamics as well as of admixture proportions, particularly of introgressed tracts along individual genomes [12].

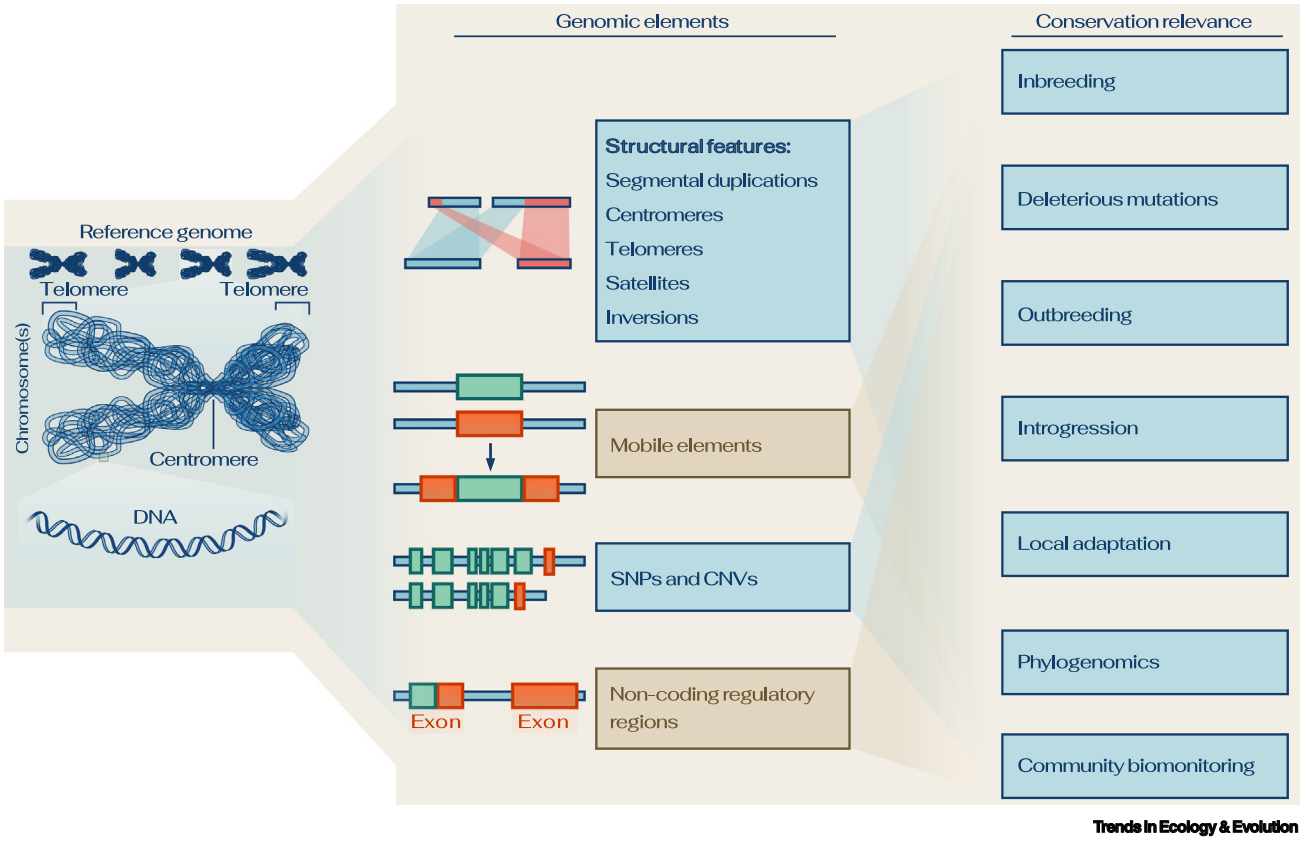


Figure 1. Reference genomes offer an (almost) complete record of the genome of a species. They characterize genomic information more thoroughly than fragmented genomes can. Importantly, they reveal structural features which often remain elusive in fragmented genome sequences. These features are relevant for conservation genomics applications. Abbreviations: CNV, copy number variants; SNP, single nucleotide polymorphism.

Local adaptation and genetic rescue

The use of reference genomes in population genomics facilitates the identification of traits under natural selection that form the basis and architecture of local adaptations, and ultimately of speciation. Reference genomes provide the functional and genomic contexts for regions influenced by selection, thereby enabling association of such loci with phenotypes important to adaptation and resilience. Identifying locally adapted variants can inform definitions of conservation units and identify optimal source populations for translocations to support **genetic rescue** [13].

Phylogenetic diversity and phylogenomics

Phylogenetic diversity is essential for ecosystem stability and resilience, and is used to delineate evolutionarily distinct components of biodiversity to guide conservation priorities [e.g., **evolutionary distinct and globally endangered (EDGE) species**] [14]. Genome-scale analyses based on hundreds or thousands of loci have become the gold standard for phylogenetic inference by capturing the evolutionary histories of the targeted taxa. Reference genomes serve as the basis for **phylogenomic** analyses because they greatly improve orthology inference at the DNA and protein levels, while also facilitating inferences based on genome organization.

Structure and function of communities

Reference genomes are particularly important in **metagenomics** and **metatranscriptomics** where total DNA, or complementary DNA (cDNA) derived from RNA, from entire communities is sequenced to understand community composition, abundance, function, and dynamics. Facilitated by the availability of reference genomes, metagenomics and metatranscriptomics have been mostly applied to microbial community samples. Eukaryotic reference genomes allow DNA/cDNA reads to be assigned to higher taxa within environmental samples, leading to a more complete characterization of communities from environmental DNA (eDNA) and RNA (eRNA). This approach represents a novel means to track changes in the composition, structure, and functioning of eukaryotic communities, and thus support the biomonitoring and management of taxonomic and functional diversity in entire ecosystems.

A collective effort to conserve biodiversity

Conservation efforts need to account for genomic diversity to optimize management strategies. Accounting for genomic diversity will aid in maintaining population viability and preserving adaptive potential to respond to environmental change. The availability of reference genomes will provide a solid, quantitative, and comparable foundation for biodiversity assessments, conservation, management, and restoration.

Author contributions

Giulio Formenti, Kathrin Theissinger, and Carlos Fernandes led the writing of the manuscript. Details on contributions to the initial discussion, to literature survey, drafting, reviewing of the manuscript and design of the figure can be found in the Supplemental information online.

Acknowledgments

We thank Fabien Condamine, Love Dalén, Richard Durbin, Bruno Fosso, Roderic Guigó, Marc Hanikenne, Alberto Pallavicini, Olga Vinnere Pettersson, Xavier Turon, and Detlef Weigel for their contributions to the manuscript, as well as the whole ERGA community for their ongoing support. We refer to the supplemental information online for acknowledgements for individual authors.

Declaration of interests

The authors declare no conflicts of interest.

Supplemental information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tree.2021.11.008.

[1] The Rockefeller University, 1230 York Ave, New York, NY 10065, USA
[2] LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Georg-Voigt-Str. 14-16, 60325 Frankfurt/Main, Germany
[3] University of Koblenz-Landau, Institute for Environmental Sciences, Fortstrasse 7, 76829 Landau, Germany
[4] Senckenberg Biodiversity and Climate Research Centre, Georg-Voigt-Str. 14-16, 60325 Frankfurt/Main, Germany
[5] CE3C - Centre for Ecology, Evolution and Environmental Changes, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[6] Faculdade de Psicologia, Universidade de Lisboa, Alameda da Universidade, 1649-013 Lisboa, Portugal
[7] University of Cambridge, Department of Genetics, Cambridge CB2 3EH, UK
[8] Wellcome Sanger Institute, Hinxton, UK
[9] Universita degli Studi di Milano, Via Celoria 26, 20133 Milano, Italy
[10] University of Göttingen, Department of Animal Evolution and Biodiversity, Untere Karspüle, 2, 37073, Germany
[11] University of Florence, Department of Biology, Via Madonna del Piano 6, Sesto Fiorentino (FI) 50019, Italy
[12] CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos Genéticos, Rua Padre Armando Quintas, 7, 4485-661, Portugal
[13] Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, Av. Américo Vespucio, 26, 41092, Spain
[14] Dept. of Ecology and Genetics, Uppsala University, Norbyvägen 18D, 75246, Sweden
[15] German Primate Center, Kellnerweg 4, 37077 Göttingen, Germany
[16] InBios - Conservation Genetics Lab, University of Liege, Chemin de la Vallée 4, 4000, Belgium
[17] Centre for Ecological and Evolutionary Synthesis, University of Oslo, Blindernveien 31, 0371 Oslo, Norway
[18] Centre for Coastal Research, University of Agder, Gimlemoen 25j, 4630 Kristiansand, Norway
[19] Groningen Institute of Evolutionary Life Sciences University of Groningen Nijenborgh, 9747, AG, Groningen, the Netherlands
[20] Center for Coastal Studies, 5 Holway Avenue, Provincetown, MA 02657, USA
[21] Marine and Freshwater Research Institute, Fornubúðir, 5, 220 Hanafjörður, Iceland
[22] Department of Biology, University of Antwerp, Groenenborgerlaan 171, 2020, Belgium
[23] University of Göttingen, Institute for Microbiology and Genetics, Dept. of Applied Bioinformatics, Goettingen Center for Molecular Biosciences (GZMB), Campus Institute Data Science (CIDAS), Goldschmidtstr. 1, 37077, Germany
[24] Institute of Zoology, University of Cologne, Zülpicherstrasse 47b, D-50674, Germany
[25] Villum Center for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Denmark, Build 3, Universitetsparken 15, Copenhagen 2100, Denmark
[26] China National Genebank, BGI-Shenzhen, Jinsha Road, Dapeng District, Shenzhen 518083, China
[27] Leibniz Institute for Zoo and Wildlife Research (IZW), Alfred-Kowalke-Str 17, 10315 Berlin, Germany
[28] Institute for Insect Biotechnology, Justus-Liebig University Gießen, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany
[29]Joint first authors.
[30]This work originates from a collective effort within the ERGA Consortium (Box 1).
[31]Consortium author information: Giulio Formenti, Kathrin Theissinger, Carlos Fernandes, Iliana Bista, Aureliano Bombarely, Christoph Bleidorn, Fedor Čiampor, Claudio Ciofi, Angelica Crottini, José A. Godoy, Jacob Hoglund, Joanna Malukiewicz, Alice Mouton, Rebekah A. Oomen, Sadye Paez, Per Palsbøll, Christophe Pampoulie, María José Ruiz-López, Hannes Svardal, Constantina Theofanopoulou, Jan de Vries, Ann-Marie Waldvogel, Goujie Zhang, Camila J. Mazzoni, Erich Jarvis, Miklós Bálint, Sargis A. Aghayan, Tyler S. Alioto, Isabel Almudi, Nadir Alvarez, Paulo C. Alves, Isabel R Amorim, Agostinho Antunes, Paula Arribas, Petr Baldrian, Paul R Berg, Giorgio Bertorelle, Astrid Böhne, Andrea Bonisoli-Alquati, Ljudevit L Boštjančić, Bastien Boussau, Catherine M Breton, Elena Buzan, Paula F Campos, Carlos Carreras, L. Fllipe Castro, Luis J. Chueca, Elena Conti, Robert Cook-Deegan, Daniel Croll, Mónica V Cunha, Frédéric Delsuc, Alice B. Dennis, Dimitar Dimitrov, Rui Faria, Adrien Favre, Olivier D. Fedrigo, Rosa Fernández, Gentile Francesco Ficetola, Jean-François Flot, Toni Gabaldón, Dolores R. Galea Agius, Guido R. Gallo, Alice M. Giani, M. Thomas P Gilbert, Tine Grebenc, Katerina Guschanski, Romain Guyot, Bernhard Hausdorf, Oliver Hawlitschek, Peter D Heintzman, Berthold Heinze, Michael Hiller, Martin Husemann, Alessio Iannucci, Iker Irisarri, Kjetill S Jakobsen, Sissel Jentoft, Peter Klinga, Agnieszka Kloch, Claudius F Kratochwil, Henrik Kusche, Kara KS Layton, Jennifer A Leonard, Emmanuelle Lerat, Gianni Liti, Tereza Manousaki, Tomas Marques-Bonet, Pável Matos-Maraví, Michael Matschiner, Florian Maumus, Ann M. Mc Cartney, Shai Meiri, José Melo-Ferreira, Ximo Mengual, Michael T. Monaghan, Matteo Montagna, Robert W Mysłajek, Marco T Neiber, Violaine Nicolas, Marta Novo, Petar Ozretić, Ferran Palero, Lucian Pârvulescu, Marta Pascual, Octávio S. Paulo, Martina Pavlek, Cinta Pegueroles, Loïc Pellissier, Graziano Pesole, Craig R Primmer, Ana Riesgo, Lukas Rüber, Diego Rubolini, Daniele Salvi, Ole Seehausen, Matthias Seidel, Simona Secomandi, Bruno Studer, Spyros Theodoridis, Marco Thines, Lara Urban, Anti Vasemägi, Adriana Vella, Noel Vella, Sonja C Vernes, Cristiano Vernesi, David R Vieites, Robert M Waterhouse, Christopher W Wheat, Gert Wörheide, Yannick Wurm, and Gabrielle Zammit.
Affiliations: Giulio Formenti, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA; Kathrin Theissinger, LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Georg-Voigt-Str. 14-16, 60325 Frankfurt/Main, Germany, University of Koblenz-Landau, Institute for Environental Sciences, Fortstrasse 7, 76829 Landau, Germany, Senckenberg Biodiversity and Climate Research Centre, Georg-Voigt-Str. 14-16, 60325 Frankfurt/Main, Germany; Carlos Fernandes, CE3C - Centre for Ecology, Evolution and Environmental Changes, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal, Faculdade de Psicologia, Universidade de Lisboa, Alameda da Universidade, 1649-013 Lisboa, Portugal; Iliana Bista, University of Cambridge, Department of Genetics, Downing Street, CB2 3EH, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK, Naturalis Biodiversity Center, Darwinweg 2, Leiden, 2333 CR, The Netherlands; Aureliano Bombarely, Universita degli Studi di Milano, Via Celoria 26, 20133 Italy; Christoph Bleidorn, University of Göttingen, Department of Animal Evolution and Biodiversity, Untere Karspüle, 2, 37073 Germany; Fedor Čiampor, Plant Science and Biodiversity Centre, Slovak Academy of Science, Dúbravská cesta 9, 845 23 Bratislava, Slovakia; Claudio Ciofi, University of Florence, Department of Biology, Via Madonna del Piano 6, Sesto Fiorentino (FI) 50019, Italy; Angelica Crottini, CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal, Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, 4099-002 Porto, Portugal, BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal; José A Godoy, Estación Biológica de Doñana, Calle Americo Vespucio 26, 41092 Spain; Jacob Hoglund, Dept. of Ecology and Genetics, Uppsala University, Norbyvägen 18D, 75246, Sweden; Joanna Malukiewicz, German Primate Center, Kellnerweg 4, 37077, Germany; Alice Mouton, InBios - Conservation Genetics Lab, University of Liege, Chemin de la Vallée 4, 4000, Belgium; Rebekah A Oomen, Centre for Ecological and Evolutionary Synthesis, University of Oslo, Blindernveien 31, 0371 Oslo, Norway, Centre for Coastal Research, University of Agder, Gimlemoen 25j, 4630 Kristiansand, Norway; Sadye Paez, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA; Per Palsbøll, Groningen Institute of Evolutionary Life Sciences University of Groningen Nijenborgh, 9747 AG Groningen, The Netherlands, Center for Coastal Studies, 5 Holway Avenue, Provincetown, MA 02657, USA; Christophe Pampoulie, Marine and Freshwater Research Institute, Fornubúðir, 5,220, Hanafjörður, Iceland; María José Ruiz-López, Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, Av. Américo Vespucio, 26, 41092, Spain, CIBER Epidemiología y Salud Pública (CIBERESP), Spain; Hannes Svardal, Department of Biology, University of Antwerp, Groenenborgerlaan 171, 2020, Belgium; Constantina Theofanopoulou, Rockefeller University, 1230 York Ave, New York, NY 10065, USA; Jan de Vries, University of Goettingen, Institute for Microbiology and Genetics, Dept. of Applied Bioinformatics, Goettingen Center for Molecular Biosciences (GZMB), Campus Institute Data Science (CIDAS), Goldschmidtstr. 1, 37077, Germany; Ann-Marie Waldvogel, Institute of Zoology, University of Cologne, Zülpicherstrasse 47b, D-50674, Germany; Goujie Zhang, Villum Center for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Build 3, Universitetsparken 15, Copenhagen, 2100, Denmark, China National Genebank, BGI-Shenzhen, Jinsha Road,

Dapeng District, Shenzhen 518083, China, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Jiaochang East Road, Wupan District, Kunming, Yunnan 650223, China; Camila J Mazzoni, Leibniz Institute for Zoo and Wildlife Research (IZW), Alfred-Kowalke-Str 17 10315 Berlin, Germany, Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Koenigin-Luise-Str 6-8, 14195 Berlin, Germany; Erich Jarvis, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA; Miklós Bálint, LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Georg-Voigt-Str. 14-16, 60325 Frankfurt/Main, Germany, Institute for Insect Biotechnology, Justus-Liebig University Gießen, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany, Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Georg-Voigt-Str. 14-16, 60325 Frankfurt/Main, Germany; Sargis A Aghayan, Chair of Zoology, Yerevan State University, 1 Alex Manoogian, Yerevan, 0025, Armenia; Tyler S Alioto, CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain; Isabel Almudi, Department of Genetics, Microbiology and Statistics and IRBio, Universitat de Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain; Nadir Alvarez, Geneva Natural History Museum, 1 route de Malagnou, 1208 Geneva, Switzerland, University of Geneva, Department of Genetics & Evolution, University of Geneva, 4 Boulevard d'Yvoy, 1205 Geneva, Switzerland; Paulo C Alves, CIBIO/InBIO, University of Porto, Campus of Vairão, 4485-661, Vila do Conde, Portugal, Dep Biology, Faculty of Sciences, University of Porto, R Campo Alegre, s/n, 4169-007 Porto, Portugal, Wildlife Biology Program, University of Montana, Missoula, MT 59812, USA; Isabel R Amorim, cE3c – Centre for Ecology, Evolution and Environmental Changes / Azorean Biodiversity Group and Universidade dos Açores, Universidade dos Açores, Rua Capitão João d'Àvila, Pico da Urze, 9700-042 Angra do Heroísmo, Azores, Portugal; Agostinho Antunes, CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, 4450-208 Porto, Portugal, Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal; Paula Arribas, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), S.C. La Laguna, 38206, Spain; Petr Baldrian, Institute of Microbiology of the Czech Academy of Sciences, Vídenská 1083, Praha 4 14220, Czech Republic; Paul R Berg, University of Agder, Centre for Coastal Research, 25, Universitetsveien, Kristiansand N-4630, Norway, Norwegian Institute for Water Research (NIVA), 21, Gaustadallèen, Oslo N-0349, Norway; Giorgio Bertorelle, Department of Life Sciences and Biotechnology, 46, via Borsari, Ferrara 44121, Italy; Astrid Böhne, Centre for Molecular Biodiversity Research, Zoological Reseach Museum Alexander Koenig Bonn, Leibniz Institute for Animal Biodiversity, Adenauerallee 160, 53113 Bonn, Germany; Andrea Bonisoli-Alquati, California State Polytechnic University, Pomona, 3801 W Temple Avenue, Pomona, CA 91767, USA; Ljudevit L Boštjančić, LOEWE Centre for Translational Biodiversity Genomics, Senckenberg Biodiversity and Climate Research Centre, Georg-Voigt-Str. 14-16, 60325 Frankfurt/Main, Germany; Bastien Boussau, UMR5558, CNRS, Université Lyon 1, Université de Lyon, Bât. Grégor Mendel, 43 bd du 11 novembre 1918, Villeurbanne, France 69622; Catherine M Breton, Alliance Bioiversity CIAT, Europe – Montpellier Office, Bioversity International France, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France, French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, F-34398 Montpellier, France; Elena Buzan, University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Glagoljaška 8, 6000 Koper, Slovenia, Environmental Protection College, Trg mladosti 7, 3320 Velenje, Slovenia; Paula F Campos, CIIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Avenida General Norton de Matos, Matosinhos 4450-208, Portugal; Carlos Carreras, Department of Genetics, Microbiology and Statistics and IRBio, Universitat de Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain; L. Fllipe Castro, Interdisciplinary Centre of Marine and Environmental Research, Avenida General Norton de Matos, S/N, Matosinhos 4450-208, Portugal, Faculty of Sciences University of Porto, Rua Campo Alegre s/n, Porto 4169-007, Portugal; Luis J Chueca, University of the Basque Country (UPV/EHU), 7, Paseo de la Universidad, Vitoria-Gasteiz 01006, Spain, LOEWE-Centre for Translational Biodiversity Genomics (LOEWE-TBG), 25, Senckenberganlage, Frankfurt am Main 60325, Germany; Elena Conti, University of Zurich, Department of Systematic and Evolutionary Botany, 107 Zollikerstrasse, Zurich 8008, Switzerland; Robert Cook-Deegan, Arizona State University, 1800 I (Eye) Street, NW, Washington, DC 20006, USA; Daniel Croll, University of Neuchatel, 11, Emile-Argand, Neuchatel CH-2000, Switzerland; Mónica V Cunha, Centre for Ecology, Evolution and Environmental Changes (CE3C), Campus da Faculdade de Ciências da Universidade de Lisboa, C2 Building, Campo Grande 1749-016, Lisboa, Portugal, BioISI- Biosystems and Integrative Sciences Institute, Campus da Faculdade de Ciências da Universidade de Lisboa, Teclabs Building, Campo Grande, 1749-016 Lisboa, Portugal; Frédéric Delsuc, Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier, France; Alice B Dennis, University of Potsdam, Institute for Biochemistry & Biology, Karl-Liebknecht-Str. 24-25, House 26, Room 2.77, 14476 Potsdam, Germany; Dimitar Dimitrov, Department of Natural History, University Museum of Bergen, University of Bergen, P.O. Box 7800, 5020 Bergen, Norway; Rui Faria, CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos; InBIO, Laboratório Associado, Universidade do Porto, Campus Agrário de Vairão, Rua Padre Armando Quintas N7, 4485-661 Vairão, Portugal; Adrien Favre, Senckenberg Research Institute and Natural History Museum, 25 Senckenberganlage, Frankfurt/Main D-60325, Germany; Olivier D Fedrigo, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA; Rosa Fernández, Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 37-49 Passeig marítim de la Barceloneta, Barcelona 08003, Spain; Gentile Francesco Ficetola, Università degli Studi di Milano, Via Celoria 10, 20133 Milano Italy, Univ. Grenoble-Alpes, F-38000 Grenoble, France; Jean-François Flot, Université libre de Bruxelles (ULB), C.P. 160/12, Avenue F. D. Roosevelt 50, 1050 Brussels, Belgium; Toni Gabaldón, Barcelona Supercomputing Centre (BSC-CNS), Jordi Girona, 29. , 08034 Barcelona, Spain, Institute for Research in Biomedicine (IRB), Carrer de Baldiri Reixac, 10, 08028 Barcelona, Spain, Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain; Dolores R Galea Agius, Centre of Molecular Medicine and Biobanking, Tal-Qroqq, Msida, MSD2080, Malta, Ġ.F. Abela Junior College, Department of Biology, Ġużè Debono Square, Msida MSD 1252, Malta; Guido R Gallo, Department of Biosciences, University of Milan, Milan, Italy, 26, Via Celoria, Milan 20133, Italy; Alice M Giani, Weill Cornell Medicine, 1300 York Ave, New York, NY 10065, USA; M. Thomas P Gilbert, Center for Evolutionary Hologenomics, The GLOBE Institute, The University of Copenhagen, 5A, Oester Farimagsgade, Copenhagen, 1353, Denmark, University Museum, NTNU, 47B, Erling Skakkes gate, Trondheim, Norway; Tine Grebenc, Slovenian Forestry Institute, Večna pot 2, SI-1000 Ljubljana, Slovenia; Katerina Guschanski, The University of Edinburgh, Institute of Evolutionary Biology, School of Biological Sciences, Ash worth Laboratories, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK, Uppsala University, Department of Ecology and Genetics/Animal Ecology, Norbyvägen 18D, SE-752 36 Uppsala, Sweden; Romain Guyot, Institut de Recherche pour le Développement, 911, ave Agropolis, Montpellier, 34394, France; Bernhard Hausdorf, Leibniz Institute for the Analysis of Biodiversity Change, Zoological Museum Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany; Oliver Hawlitschek, Leibniz Institute for the Analysis of Biodiversity Change, Zoological Museum Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany; Peter D Heintzman, The Arctic University Museum of Norway, UiT - The Arctic University of Norway, P. O. Box 6050, Langnes, Tromsø, N-9037, Norway; Berthold Heinze, Austrian Research Centre for Forests (BFW), 8, Seckendorff-Gudent Weg, Vienna, 1130, Austria; Michael Hiller, LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt, Germany, Senckenberg Society for Nature Research, Senckenberganlage 25, 60325 Frankfurt, Germany, Goethe-University, Faculty of Biosciences, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany; Martin Husemann, Leibniz Institute for the Analysis of Biodiversity Change, Zoological Museum Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany; Alessio Iannucci, University of Florence, Department of Biology, Via Madonna del Piano 6, Sesto Fiorentino (FI) 50019, Italy; Iker Irisarri, University of Goettiingen, 1 Goldschmidtstr., Goettiingen, 37077, Germany, Campus Institute Data Science (CIDAS), Goettiingen, Germany; Kjetill S Jakobsen, Centre for Ecological and Evolutionary Synthesis (CEES), Dept. of Biosciences, University of Oslo, PO Box 1066 Blindern, NO-0316 Oslo, Norway; Sissel Jentoft, University of Oslo, Department of Biosciences, Centre for Ecological and Evolutionary Synthesis (CEES), Blindernveien 31, 0371 Oslo, Norway; Peter Klinga, Technical University in Zvolen, Faculty of Forestry, Department of Phytology, 24, T.G. Masaryka, Zvolen, 960 01, Slovak Republic, DIANA – Carpathian Wildlife Research, 47, Mládežnícka, Banská Bystrica 974 04, Slovak Republic; Agnieszka Kloch, Institute of Functional Biology and Ecology, Department of Ecology, Univeristy of Warsaw, Faculty of Biology, Iii, Miecznikowa 1, 02-096 Warszawa, Poland; Claudius F Kratochwil, Institute of Biotechnology, HiLIFE, University of Helsinki, 1, Viikinkaari, Helsinki, 00790, Finland; Henrik Kusche, Leibniz Institute for the Analysis of Biodiversity Change, Zoological Museum Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany; Kara KS Layton, University of Aberdeen, Zoology Building, Tillydrone Ave, University of Aberdeen, Aberdeen, AB24 2TZ, UK; Jennifer A Leonard, Estación Biológica de Doñana (EBD-CSIC), Avd. Americo Vespucio 26, Seville, 41092, Spain; Emmanuelle Lerat, Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, Bat. Mendel, 43 bd du 11 novembre 1918, 69622 Villeurbanne cedex, France; Gianni Liti, Université Côte d'Azur, CNRS, INSERM, IRCAN, 28 Avenue de Valombrose, 06107 NICE Cedex 2, France; Tereza Manousaki, Hellenic Centre for Marine Research (HCMR), Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Former U. S. Base of Gournes, P.O. Box 2214, 71003, Heraklion, Crete, Greece; Tomas Marques-Bonet, Institute of Evolutionary Biology (UPF-CSIC), PRBB, Dr. Aiguader 88, 08003, Spain, Catalan Institution of Research and Advanced Studies (ICREA), Passeig de Lluís Companys, 23, 08010 Barcelona, Spain, CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain; Pável Matos-Maraví, Biology Centre of the Czech Academy of Sciences, Institute of Entomology, Branišovská 1160/31, 37005 České Budějovice, Czech Republic; Michael Matschiner, University of Oslo, 1, Sars' gate, Oslo, 0562, Norway; Florian Maumus, Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France; Ann M Mc Cartney, National Institute of Health, 49 Convent Drive, Bethesda, MD 20892, USA; Shai Meiri, Tel Aviv University, School of Zoology, the Steinhardt Museum of Natural History, 12 Klausner street, Tel Aviv 6997801, Israel; José Melo-Ferreira, CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Universidade do Porto, 7, Rua Padre Armando Quintas, Vairão, 4485-661, Portugal, Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, s/n, Rua do Campo Alegre, Porto 4169-007, Portugal; Ximo Mengual, Centre of Taxonomy and Evolutionary Research, Zoological Reseach Museum Alexander Koenig Bonn, Leibniz Institute for Animal Biodiversity, Adenauerallee 160, 53113 Bonn, Germany; Michael T Monaghan, Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), 301, Müggelseedamm, Berlin, 12587, Germany, Institut für

Biologie, Freie Universität Berlin, 1-3, Königin-Luise-Str., Berlin 12489, Germany; Matteo Montagna, Dipartimento di Scienze Agrarie e Ambientali - Università degli Studi di Milano, 2, Via Celoria, Milan, I-20133, Italy, BAT Center - Interuniversity Center for Studies on Bioinspired Agro-Environmental Technology, University of Napoli Federico II, 100, Via Università, Portici I-80055, Italy; Robert W Mysłajek, University of Warsaw, Faculty of Biology, Institute of Functional Biology and Ecology, Department of Ecology, Biological and Chemical Research Centre, 101 Zwirki i Wigury, Warszawa, 02-089, Poland; Marco T Neiber, Leibniz Institute for the Analysis of Biodiversity Change, Zoological Museum Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany; Violaine Nicolas, Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, CP51, 57 rue Cuvier, 75005 Paris, France; Marta Novo, Complutense University of Madrid, José Antonio Nováis, 12, Madrid, 28040, Spain; Petar Ozretić, Ruđer Bošković Institute, Bijenička cesta 54, Zagreb, 10000, Croatia; Ferran Palero, Institut Cavanilles de Biodiversitat i Biologia Evolutiva (ICBIBE), Universitat de Valencia, Carrer del Catedrátic José Beltrán Martinez 2, 46980 Paterna, Spain; Lucian Pârvulescu, Department of Biology-Chemistry, Faculty of Chemistry, Biology, Geography, West University of Timisoara, 16A Pestalozzi St., Timisoara, 300115, Romania, Environmental Advanced Research Institute, West University of Timisoara, Vasile Pârvan 4 Bd., Timisoara 300223, Romania; Marta Pascual, Department of Genetics, Microbiology and Statistics and IRBio, Universitat de Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain; Octávio S Paulo, cE3c – Centre for Ecology, Evolution and Environmental Changes, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, P-1749-016, Lisboa; Martina Pavlek, Ruđer Bošković Institute, Bijenička cesta 54, Zagreb, 10000, Croatia; Cinta Pegueroles, Department of Genetics, Microbiology and Statistics and IRBio, Universitat de Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain; Loïc Pellissier, ETH Zürich, Landscape Ecology, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland, Swiss Federal Institute for Forest Snow and Landscape Research WSL, Unit of Land Change Science, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland; Graziano Pesole, Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari "A. Moro", Campus "E. Quagliariello", voa Orabona, 4, 70126 Bari, Italy, Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Consiglio Nazionale delle Ricerche, Campus "E. Quagliariello", voa Orabona, 4, 70126 Bari, Italy, Consorzio Interuniversitario Bioteconologie, via Flavia, 23/1, 34148 Trieste, Italy; Craig R Primmer, Organismal and Evolutionary Biology Research Program, University of Helsinki, PO Box 56, 00014, University of Helsinki, Finland, Institute of Biotechnology (HiLIFE), University of Helsinki, PO Box 56, 00014, Finland; Ana Riesgo, Museo Nacional de Ciencias Naturales – CSIC, C/ José Gutiérrez Abascal, 2, Madrid, 28028, Spain, Natural History Museum of London, Cromwell Road, SW7 5BD London, UK; Lukas Rüber, Naturhistorisches Museum Bern, 15 Bernastrasse, Bern, 3005, Switzerland, Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, 6 Balzerstrasse, Bern 3012, Switzerland; Diego Rubolini, Dipartimento di Scienze e Politiche Ambientali, Università degli Studi di Milano, Via Celoria 26, Milano, I-20133, Italy; Daniele Salvi, Department of Health, Life & Environmental Sciences - University of L'Aquila, Via Vetoio snc, 67100 L'Aquila-Coppito, Italy; Ole Seehausen, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland, Eawag Swiss Federal Institute for Aquatic Science & Technology, Seestrasse 79, 6047 Kastanienbaum, Switzerland; Matthias Seidel, Leibniz Institute for the Analysis of Biodiversity Change, Zoological Museum Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany; Simona Secomandi, Department of Biosciences, University of Milan, Milan, Italy, 26, Via Celoria, Milan 20133, Italy; Bruno Studer, Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Universitaetstrasse 2, 8092 Zurich, Switzerland; Spyros Theodoridis, Senckenberg Biodiversität und Klima Forschungszentrum: Frankfurt, DE, 25 Senckenberganlage, Frankfurt am Main 60325 Germany; Marco Thines, Goethe University, Institute of Ecology, Evolution and Diversity, Max-von-Laue-Str. 13, Frankfurt am Main, 60438, Germany, Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Senckenberganlage 25, Frankfurt am Main 60325, Germany, LOEWE Translational Biodiversity Genomics, Georg-Voigt-Str. 14-16, Frankfurt am Main 60325, Germany; Lara Urban, Department of Anatomy, University of Otago, 270 Gt King Street, Dunedin, 9016, New Zealand; Anti Vasemägi, Department of Aquatic Resources, Swedish University of Agricultural Sciences, 2, Stångholmsvägen, Drottningholm, 178 93, Sweden, Chair of Aquaculture, Institute of Veterinary Medicine and Animal, Estonian University of Life Sciences, 56A, Kreutzwaldi, Tartu 51006, Estonia; Adriana Vella, Conservation Biology Research Group, University of Malta, Taq-Qroqq, Msida, MSD2080, Malta; Noel Vella, Conservation Biology Research Group, University of Malta, Msida, MSD2080, Malta; Sonja C Vernes, School of Biology, The University of St Andrews, St Andrews, Fife, UK, Max Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen, the Netherlands; Cristiano Vernesi, Forest Ecology, Research and Innovation Centre-Fondazione Edmund Mach, 1, via Edmund Mach, San Michele all'Adige, 38010, Italy; David R Vieites, Museo Nacional de Ciencias Naturales -CSIC, Calle José Gutierrez Abascal 2, 28006, Madrid, Spain; Robert M Waterhouse, Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; Christopher W Wheat, Department of Zoology, Stockholm University, 18B, Svante Arrheniusväg, Stockholm, S-10691, Sweden; Gert Wörheide, Ludwig-Maximilians-Universität München, Dept. of Earth and Environmental Sciences, Paleontology & Geobiology, Richard-Wagner-Str. 10, Munich, 80333, Germany, Ludwig-Maximilians-Universität München, GeoBio-Center, Richard-Wagner-Str. 10, Munich 80333, Germany, SNSB - Bavarian State Collections for Palaeontology and Geology, Richard-Wagner-Str. 10, Munich 80333, Germany; Yannick Wurm, Queen Mary University of London, Mile End Road, London, E5 8HS, United Kingdom, Alan Turing Institute, London NW1 2DB, UK; Gabrielle Zammit, Department of Biology, University of Malta, Biomedical Sciences Building, MSD2080, Msida, Malta, Centre for Molecular Medicine and Biobanking, Biomedical Sciences Building, MSD2080 Msida, Malta

*Correspondence:
miklos.balint@senckenberg.de (M. Bálint).

https://doi.org/10.1016/j.tree.2021.11.008

References

1. Shaw, F. *et al.* (2020) COPO: a metadata platform for brokering FAIR data in the life sciences. *F1000Res* 9, 495
2. Rhie, A. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592, 737–746
3. Howe, K.L. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.* 49, D884–D891
4. Brandies, P. *et al.* (2019) The value of reference genomes in the conservation of threatened species. *Genes (Basel)* 10, 846
5. Ballouz, S. *et al.* (2019) Is it time to change the reference genome? *Genome Biol.* 20, 159
6. Valiente-Mullor, C. *et al.* (2021) One is not enough: on the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS Comput. Biol.* 17, e1008678
7. Llamas, B. *et al.* (2019) A strategy for building and using a human reference pangenome. *F1000Res* 8, 1751
8. Mérot, C. *et al.* (2020) A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol. Evol.* 35, 561–572
9. Kardos, M. *et al.* (2016) Genomics advances the study of inbreeding depression in the wild. *Evol. Appl.* 9, 1205–1218
10. Dussex, N. *et al.* (2021) Population genomics of the critically endangered kākāpō. *Cell Genomics* 1, 100002
11. Leitwein, M. *et al.* (2021) Associative overdominance and negative epistasis shape genome-wide ancestry landscape in supplemented fish populations. *Genes (Basel)* 12, 524
12. Rogers, J. *et al.* (2019) The comparative genomics and complex population history of *Papio* baboons. *Sci. Adv.* 5, eaau6947
13. Flanagan, S.P. *et al.* (2018) Guidelines for planning genomic assessment and monitoring of locally adaptive variation to inform species conservation. *Evol. Appl.* 11, 1035–1052
14. Owen, N.R. *et al.* (2019) Global conservation of phylogenetic diversity captures more than just functional diversity. *Nat. Commun.* 10, 859

SUMMARY AND CONCLUDING REMARKS

In this thesis work, I focused on bird genomics from field work, to laboratory, to bioinformatics. I exploited the potential of newest sequencing technologies and their combination for *de novo* genome assembly, with the final aim of reconstructing the complete chromosomes of each species. Moreover, I was able to experience the importance of genomic resources in deciphering the biology and evolution of bird species, also evaluating their relevance in the management of threatened species. The thesis main focus is on the reference genome of the barn swallow (*Hirundo rustica*), an iconic migratory passerine bird with a close association with humans. The nuclear genome section was also accompanied by a mitochondrial DNA study, testing the value of complete mitogenomes assemblies in resolving the phylogeographic history of the species and the relationships between subspecies.

With my work, I outlined the importance to sequence the entire genomic sequence of species, both nuclear and mitochondrial. Recent technological advances, improvements in scalability, quality of sequencing technologies and computational resources, together with the establishment of international genome sequencing initiatives, allowed the generation of cost-effective genome assemblies at the chromosome-level also for non-model species. **Chapter 1** outlined the contribution of different sequencing technologies in the assembly process, presenting the assembly pipelines used by the Vertebrate Genomes Project for the generation of chromosome-level assemblies for all vertebrates, using both short- and long-read sequencing technologies[48]. Long noisy reads such as PacBio CLR are not enough to assemble both contiguous and accurate genome assemblies. Indeed, a base calling accuracy of 99.99% could be only reached with their combination with highly-accurate short reads. Moreover, data that add information about chromosomal conformation and physical distances between known DNA motifs in the sequence, were found pivotal to assemble the entire chromosomes of a species. With the VGP pipelines it was also possible to assess the impact of repetitive regions in genome assembly, assemble the complex sex chromosomes, resolve the retention of false duplications, correctly phase divergent haplotypes and outline the importance of manual curation at the end of automated pipelines. Moreover, it was also evaluated the potential of these new chromosome-level assemblies in answering biological questions. Finally, transcriptomic data (short-read RNASeq, Iso-Seq and Nanopore long-reads) were used for the functional annotation of the genomes, providing a comparable set of genes among species that will facilitate comparative studies.

The assemblies I generated with the VGP pipelines, or I contributed to, are publicly available on NCBI (**Table 1**). All the assembled genomes fulfilled the VGP standard metrics in terms of contiguity and chromosome assignment. In **chapter 2**, **4** and **5** I extensively presented the new chromosome

level assemblies for the barn swallow[401], the lesser kestrel (Bounas et al., forthcoming), and the European nightjar[402], respectively, which exceeded all the threshold statistics set for VGP-quality assemblies.

Species	Common name	Primary/alternate NCBI a.n.	Length (Gbp)	Scaffold N50 (Mbp)	n° chr.	% chr.
Tauraco erythrolophus	Red-crested turaco	GCA_009769465.1 GCA_009764505.1	1.25	85.6	31+ZW	96.2
Geothlypis trichas	Common yellowthroat	GCA_009764595.1 GCA_009764545.1	1.08	72.5	32+ZW	98.67
Phoenicopterus ruber ruber	American flamingo	GCA_009819775.1 GCA_009819805.1	1.25	85.5	31+ZW	99.23
Acanthisitta chloris	Rifleman	GCA_016904835.1 GCA_016880875.1	1.08	40.8	36+ZW	99.04
Hirundo rustica	Barn swallow	GCA_015227805.3 GCA_015227815.3	1.11	76.2	39+ZW	98.20
Caprimulgus europaeus	European nightjar	GCA_907165065.1 GCA_907165095.1	1.18	82.6	35+ZW	99.30
Falco naumanni	Lesser kestrel	GCF_017639655.2 GCA_017639645.1	1.22	91.8	22+ZW	98.92

Table 1. Assemblies available on NCBI. NCBI accession numbers (a.n.) for both the primary pseudo-haplotype and alternate haplotype, genome length in Gbp, scaffold N50 in Mbp, chromosomes number (n° chr.) and the percentage of the genomic sequences assigned to chromosomes (% chr.) for the primary assembly are reported for all species. For the lesser kestrel, which is a trio assembly, the a.n. for the paternal and maternal haplotypes are reported, with statistics and chromosomes number for the paternal assembly integrated with the W female sex chromosome, which was chosen as the representative genome for the species.

The public availability of well-annotated high-quality reference genomes spanning the entire tree of life is important for biological and evolutionary studies. Reference genomes will help in addressing specific biological questions, deciphering the molecular bases of many phenotypic traits to truly understand the complexity of life. Reference genomes with a comparable quality and functional completeness will further facilitate studies of orthologous genes evolution, chromosomal evolution, pathogens and vectors, the evolution of taxonomic innovations and the reconstruction of ancestral genomes[403]. For example, the lesser kestrel reference genome will be a fundamental resource for an ongoing study that involves the assessment of how this species coped with climatic fluctuations in the past and how it is expected to cope with them in the future under the current scenario of climate change. Moreover, the European nightjar genome will help to deepen the knowledge on the biology of this cryptic bird, also boosting the sequencing of other members of the Caprimulgidae family to reconstruct a comprehensive phylogeny.

In **Chapter 2** I explored the potential of a chromosome-level assembly in gaining insight into the evolution of the barn swallow. In the study, I included the barn swallow reference genome in a comparative genomic study with other chromosome-level bird genomes and we performed a population genomics study to generate a comprehensive catalog of genetic variants for the species. We also scanned the genome to detect Linkage Disequilibrium (LD) blocks, indicative of potential signatures of selection. Comparative genomics and population genomics analyses both permitted the detection of putative conserved candidate genes under negative selection. The two top candidate genes, *camk2n2* and *bdnf*, are involved in the regulation of the glutamate signaling, and in that of neural crest cells development, respectively, two main mechanisms that guide morphological and behavioral changes under domestication[281,282]. We therefore speculated that the mechanisms underlying the strict association with humans of the barn swallow may be linked to attenuated fear response and increased tameness, which are typically under selection in domesticated species. A chromosome-level genome, together with an accurate gene annotation such as the barn swallow one therefore provides a powerful resource to study the evolution of specific traits in a species. Further analyses will be required to confirm the conservation of the two candidate genes. I also used the barn swallow reference genome to generate of the first pangenome graph of the species, including also the primary and alternate haplotypes of 5 barn swallow individuals sequenced with HiFi reads. This provided preliminary insights into the sequences and genes shared between all the individuals (core genes/genome) and the putative accessory ones. Further analysis on the pangenome, with the addition of genomes from more individuals, will be necessary to validate the putative loss of genes in some individuals, and to rule out any technical limitation of HiFi reads[404]. We also evaluated the potential of HiFi reads for *de novo* genome assembly, which permitted to quickly and easily assemble phased contig-level contiguous genomes. HiFi reads are also being implemented in the new VGP assembly pipeline v2.0, replacing PacBio CLR long-reads, and achieving a higher level of base call accuracy and greater contiguity. This eliminates the need for short-reads technologies, such as Illumina, to increase assembly accuracy. The HiFi preliminary assembly is then scaffolded only with Bionano optical maps and Hi-C data. In **Chapter 2**, we generated the catalog of genetic markers using publicly available data and our HiFi data. In particular, the HiFi reads were also found very effective in detecting genomic variation even when the samples are limited and the coverage is low. HiFi reads were also extremely effective in phasing genomic variants into haplotype blocks.

Chapter 3 reports a companion study on complete barn swallows mtDNA sequences resulted from a collaboration with Antonio Torroni's group at the University of Pavia (Italy), that complements the nuclear genome paper presented in **Chapter 2**. Complete mitogenome data allowed to build a detailed phylogeny for the species, to determine its coalescence time as well as the ages of its haplogroups,

and to better define the matrilineal relationships between subspecies. This study showed that the information contained in mitochondrial DNA is phylogenetically best exploited when the sequence variation of the entire mitogenome is available and the sequencing survey is carried out on numerous specimens samples throughout the species distribution range. Moreover, the study outlined the strong link of this widespread species with climate fluctuations and human activities, making the barn swallow an excellent indicator for monitoring and assessing the impact of current global changes on wildlife.

Finally, in **Chapter 6**, reference genomes are described as valuable tools for conservation genomics. We are currently in the middle of the sixth mass extinction[260] and advances in genome sequencing allowed the rise of the conservation genomics era. Chromosome-level reference genomes, such as those presented in this thesis work, could provide a full spectrum of a species genomic diversity, but also population genomics tools to quantify inbreeding and deleterious mutations, outbreeding and hybridization, and local adaptation and genetic rescue. Moreover, phylogenetic, phylogenomics, metagenomics and metatranscriptomics studies will benefit from these state-of-the-art genomic resources.

Future perspectives

Following the results from this thesis work, a telomere-to-telomere (T2T) assembly for the barn swallow will be generated to further improve the one presented here. The species pangenome will be complemented with a barn swallow genome for each subspecies assembled from newly HiFi sequenced data, which will better represent the entire genetic diversity of the species[231]. Additional barn swallow individuals (*H. r. rustica*), are currently being sequenced with a combination of Illumina short reads and HiFi reads (the mixed sequencing approach[405]) at the coverage we evaluated as optimal for having a reliable number of variants[401]. This combination of data will be used to perform GWAS with the final aim of identifying the genetic basis of complex traits in the barn swallow, such as those that determine individuals' fitness. The lesser kestrel assembly, instead, will be involved in a study which aims to understand how individuals coped with past climate changes, making also predictions about the effects of future ones (Bounas et al., forthcoming).

In conclusion, in this thesis work I outlined how the availability of complete reference genomes is pivotal to decipher the biology and evolution of a species, also providing reliable resources to correctly plan conservation actions of threatened species. This thesis work represents a valuable contribution in characterizing and understanding the genetic diversity of our planet. The resources and results presented here will lay the foundations for future genomic studies on the sequenced species and hopefully boost similar studies to be carried out on other bird species.

ACKNOWLEDGMENTS

I would like to thank my tutors at the University of Milan (Italy) and the Rockefeller University (New York City, NY, USA) and all my collaborators and co-authors that permitted the accomplishment of this thesis work.

Luca Gianfranceschi
Guido Gallo
Marcella Sozzoni
Elena Galati

Nicola Saino
Roberto Ambrosini
Diego Rubolini
Joan Ferrer-Obiol
Alessandra Costanzo
Manuela Caprioli

Giulio Formenti
Erich D. Jarvis
Olivier Fedrigo
Jennifer Balacco
Linelle Abueg
and all the VGL

Antonio Torroni
Anna Olivieri
Luca Ferretti
Gianluca Lombardo

Claudio Ciofi
Alessio Iannucci

Andrea Bonisoli-Alquati

Anastasios Bounas

Kerstin Howe
Joanna Collins
William Chow
James Torrance

Françoise Thibaud-Nissen

Kristina Weber
David Stucki

Arang Rhie
Adam M. Phillippy

Arkarachai Fungtammasan

I would also like to thank Prof. Erich Jarvis and Dr. Olivier Fedrigo for having welcomed me in the Vertebrate Genomes Lab at the Rockefeller University in October and November 2021. Finally, I would like to thank all the VGP, DToL and ERGA members.

APPENDIX - Additional publications

1. Costanzo, A., … **Secomandi, S.**, et al. Telomere shortening is associated with corticosterone stress response in adult barn swallows. *Curr. Zool.* **68**, (2021)

2. Parolini, M., … **Secomandi, S.**, et al. Prenatal independent and combined effects of yolk vitamin E and corticosterone on embryo growth and oxidative status in the yellow-legged gull. *J. Exp. Biol*. **222**, (2019).

3. Possenti, C. D., **Secomandi, S.**, et al. Independent and combined effects of egg pro- and anti-oxidants on gull chick phenotype. *J. Exp. Biol*. **221**, (2018).

REFERENCES

1. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
2. F. Sanger, E. O. P. T. The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochem. J* **53**, 353 (1953).
3. Sanger, F. & Thompson, E. O. P. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochem. J* **53**, 366–374 (1953).
4. Holley, R. W. *et al.* Structure of a ribonucleic acid. *Science* **147**, 1462–1465 (1965).
5. Wu, R. & Kaiser, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* **35**, 523–537 (1968).
6. Brownlee, G. G., Sanger, F. & Barrell, B. G. The sequence of 5 s ribosomal ribonucleic acid. *J. Mol. Biol.* **34**, 379–412 (1968).
7. Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237**, 82–88 (1972).
8. Gilbert, W. & Maxam, A. The nucleotide sequence of the lac operator. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3581–3584 (1973).
9. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448 (1975).
10. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507 (1976).
11. Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* **18**, 9–19 (2020).
12. Sanger, F. *et al.* Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265**, 687–695 (1977).
13. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
14. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
15. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564 (1977).

16. Maniatis, T., Jeffrey, A. & van deSande, H. Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. *Biochemistry* **14**, 3787–3794 (1975).
17. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6**, 2601–2610 (1979).
18. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* **162**, 729–773 (1982).
19. Baer, R. *et al.* DNA sequence and expression of the B95-8 Epstein—Barr virus genome. *Nature* **310**, 207–211 (1984).
20. Mullis, K. B. The unusual origin of the polymerase chain reaction. *Sci. Am.* **262**, 56–61, 64–5 (1990).
21. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).
22. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341 (1987).
23. Tabor, S. & Richardson, C. C. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proceedings of the National Academy of Sciences* vol. 84 4767–4771 (1987).
24. Murray, V. Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucleic Acids Res.* **17**, 8889 (1989).
25. Craxton, M. Linear amplification sequencing, a powerful method for sequencing DNA. *Methods* **3**, 20–26 (1991).
26. Hyman, E. D. A new method of sequencing DNA. *Anal. Biochem.* **174**, 423–436 (1988).
27. Nyrén, P., Pettersson, B. & Uhlén, M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem.* **208**, 171–175 (1993).
28. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).
29. Nyrén, P. Enzymatic method for continuous monitoring of DNA polymerase activity. *Anal. Biochem.* **167**, 235–238 (1987).
30. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363, 365 (1998).
31. GenBank overview. https://www.ncbi.nlm.nih.gov/genbank/.

32. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

33. Head, S. R. *et al.* Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* **56**, 61–4, 66, 68, passim (2014).

34. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

35. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* **55**, 641–658 (2009).

36. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* **34**, e22 (2006).

37. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

38. Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A.-P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.* **36**, e25 (2008).

39. Ruparel, H. *et al.* Design and synthesis of a 3′-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5932–5937 (2005).

40. Seo, T. S. *et al.* Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proceedings of the National Academy of Sciences* vol. 102 5926–5931 (2005).

41. Barnes, C., Balasubramanian, S., Liu, X. & Swerdlow, H. Labelled nucleotides. *Patent US7057026* (2006).

42. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).

43. Shendure, J. *et al.* Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* vol. 309 1728–1732 (2005).

44. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).

45. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).

46. Song, L. *et al.* Comparison of error correction algorithms for Ion Torrent PGM data: application to hepatitis B virus. *Sci. Rep.* **7**, 8106 (2017).

47. Adewale, B. A. Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *S. Afr. J. Lab. Clin. Med.* **9**, 1340 (2020).

48. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).

49. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

50. Küpper, C. *et al.* A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48**, 79–83 (2016).

51. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).

52. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).

53. English, A. C., Salerno, W. J. & Reid, J. G. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* **15**, 180 (2014).

54. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).

55. Tattini, L., D'Aurizio, R. & Magi, A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology* vol. 3 (2015).

56. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28**, 2711–2718 (2012).

57. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).

58. Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* **29**, 635–645 (2019).

59. Belton, J.-M. *et al.* Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).

60. Yardımcı, G. G. & Noble, W. S. Software tools for visualizing Hi-C data. *Genome Biol.* **18**, 26 (2017).

61. Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, e157 (2005).

62. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

63. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).

64. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* **46**, 2159–2168 (2018).

65. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. *Genome Biol.* **14**, 405 (2013).

66. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).

67. Merker, J. D. *et al.* Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* **20**, 159–163 (2018).

68. Zhang, H., Jain, C. & Aluru, S. A comprehensive evaluation of long read error correction methods. *BMC Genomics* **21**, 889 (2020).

69. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

70. Henry, J. P., Chich, J. F., Goldschmidt, D. & Thieffry, M. Blockade of a mitochondrial cationic channel by an addressing peptide: an electrophysiological study. *J. Membr. Biol.* **112**, 139–147 (1989).

71. Brundage, L., Hendrick, J. P., Schiebel, E., Driessen, A. J. & Wickner, W. The purified *E. coli* integral membrane protein SecY/E is sufficient for reconstitution of SecA-dependent precursor protein translocation. *Cell* **62**, 649–657 (1990).

72. Akimaru, J., Matsuyama, S., Tokuda, H. & Mizushima, S. Reconstitution of a protein translocation system containing purified SecY, SecE, and SecA from *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 6545–6549 (1991).

73. Görlich, D. & Rapoport, T. A. Protein translocation into proteoliposomes reconstituted from purified components of the endoplasmic reticulum membrane. *Cell* **75**, 615–630 (1993).

74. Simon, S. M. & Blobel, G. A protein-conducting channel in the endoplasmic reticulum. *Cell* **65**, 371–380 (1991).

75. Bustamante, J. O., Hanover, J. A. & Liepins, A. The ion channel behavior of the nuclear pore complex. *J. Membr. Biol.* **146**, 239–251 (1995).

76. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).

77. Bharagava, R. N., Purchase, D., Saxena, G. & Mulla, S. I. Chapter 26 - Applications of Metagenomics in Microbial Bioremediation of Pollutants: From Genomics to Environmental Cleanup. in *Microbial Diversity in the Genomic Era* (eds. Das, S. & Dash, H. R.) 459–477 (Academic Press, 2019). doi:10.1016/B978-0-12-814849-5.00026-5.

78. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13770–13773 (1996).

79. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).

80. Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411–413 (2017).

81. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).

82. Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193–2198 (2019).

83. Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* **25**, 1750–1756 (2015).

84. He, M., Chi, X. & Ren, J. Applications of Oxford Nanopore Sequencing in Schizosaccharomyces pombe. *Methods Mol. Biol.* **2196**, 97–116 (2021).

85. Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159 (2010).

86. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).

87. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

88. Korlach, J. Understanding accuracy in SMRT sequencing. *Pac Biosci* 1–9 (2013).

89. Loomis, E. W. *et al.* Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* **23**, 121–128 (2013).

90. Yuan, Y., Chung, C. Y.-L. & Chan, T.-F. Advances in optical mapping for genomic research. *Comput. Struct. Biotechnol. J.* **18**, 2051–2062 (2020).

91. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).

92. Savara, J., Novosád, T., Gajdoš, P. & Kriegová, E. Comparison of structural variants detected by optical mapping with long-read next-generation sequencing. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab359.

93. Cao, H. *et al.* Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience* vol. 3 (2014).

94. Shelton, J. M. *et al.* Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics* **16**, 734 (2015).

95. Aston, C., Mishra, B. & Schwartz, D. C. Optical mapping and its potential for large-scale sequencing projects. *Trends Biotechnol.* **17**, 297–302 (1999).

96. Rice, E. S. & Green, R. E. New Approaches for Genome Assembly and Scaffolding. *Annu Rev Anim Biosci* **7**, 17–40 (2019).

97. Goffeau, A. *et al.* Life with 6000 Genes. *Science* vol. 274 546–567 (1996).

98. C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).

99. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).

100. Initiative, T. A. G. & The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* vol. 408 796–815 (2000).

101. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

102. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

103. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).

104. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).

105. 10K Community of Scientists, G. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J. Hered.* (2009).

106. Meulemans, D. & Bronner-Fraser, M. Amphioxus and lamprey AP-2 genes: implications for neural crest evolution and migration patterns. *Development* **129**, 4953–4962 (2002).

107. Baker, C. V. H. The evolution and elaboration of vertebrate neural crest cells. *Curr. Opin. Genet. Dev.* **18**, 536–543 (2008).

108. Shimeld, S. M. & Holland, P. W. Vertebrate innovations. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 4449–4452 (2000).

109. Brito, P. H. & Edwards, S. V. Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* **135**, 439–455 (2009).

110. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).

111. Romanov, M. N. *et al.* The value of avian genomics to the conservation of wildlife. *BMC Genomics* **10 Suppl 2**, S10 (2009).

112. Zhang, G. *et al.* Genomics: Bird sequencing project takes off. *Nature* **522**, 34 (2015).

113. Gill, F., Donsker, D. & Rasmussen, P. IOC World Bird List 12.1. *IOC World Bird List Datasets* (2022) doi:10.14344/IOC.ML.12.1.

114. Alexander, D. J. A review of avian influenza in different bird species. *Vet. Microbiol.* **74**, 3–13 (2000).

115. Vázquez, D. P. *Auk* **115**, 1089–1091 (1998).

116. Holt, B. G. *et al.* An update of Wallace's zoogeographic regions of the world. *Science* **339**, 74–78 (2013).

117. Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Mooers, A. O. The global diversity of birds in space and time. *Nature* **491**, 444–448 (2012).

118. Bravo, G. A., Schmitt, C. J. & Edwards, S. V. What Have We Learned from the First 500 Avian Genomes? *Annu. Rev. Ecol. Evol. Syst.* **52**, 611–639 (2021).

119. Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C. & Volff, J.-N. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res.* **16**, 203–215 (2008).

120. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).

121. de Oliveira, E. H. C. *et al.* Chromosome reshuffling in birds of prey: the karyotype of the world's largest eagle (Harpy eagle, *Harpia harpyja*) compared to that of the chicken (*Gallus gallus*). *Chromosoma* **114**, 338–343 (2005).

122. Nanda, I. *et al.* Extensive gross genomic rearrangements between chicken and Old World vultures (Falconiformes: Accipitridae). *Cytogenet. Genome Res.* **112**, 286–295 (2006).

123. Joseph, S. *et al.* Chromosome Level Genome Assembly and Comparative Genomics between Three Falcon Species Reveals an Unusual Pattern of Genome Organisation. *Diversity* **10**, 113 (2018).

124. Christidis, L., Shaw, D. D. & Schodde, R. Chromosomal evolution in parrots, lorikeets and cockatoos (Aves: Psittaciformes). *Hereditas* **114**, 47–56 (2008).

125. Dussex, N. *et al.* Population genomics of the critically endangered kākāpō. *Cell Genomics* **1**, 100002 (2021).

126. Nie, W. *et al.* Avian comparative genomics: reciprocal chromosome painting between domestic chicken (*Gallus gallus*) and the stone curlew (*Burhinus oedicnemus*, Charadriiformes)--an atypical species with low diploid number. *Chromosome Res.* **17**, 99–113 (2009).

127. Burt, D. W. Origin and evolution of avian microchromosomes. *Cytogenet. Genome Res.* **96**, 97–112 (2002).

128. Axelsson, E., Webster, M. T., Smith, N. G. C., Burt, D. W. & Ellegren, H. Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Res.* **15**, 120–125 (2005).

129. Ellegren, H. Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol. Evol.* **25**, 283–291 (2010).

130. Kraus, R. H. S. & Wink, M. Avian genomics: fledging into the wild! *J. Ornithol.* **156**, 851–865 (2015).

131. Kessler, L. G. & Avise, J. C. Systematic relationships among waterfowl (Anatidae) inferred from restriction endonuclease analysis of mitochondrial DNA. *Syst. Biol.* (1984).

132. Kessler, L. G. & Avise, J. C. A comparative description of mitochondrial DNA differentiation in selected avian and other vertebrate genera. *Mol. Biol. Evol.* **2**, 109–125 (1985).

133. Burt, D. W. *et al.* The dynamics of chromosome evolution in birds and mammals. *Nature* **402**, 411–413 (1999).

134. Shetty, S., Griffin, D. K. & Graves, J. A. Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. *Chromosome Res.* **7**, 289–295 (1999).

135. Guttenbach, M. *et al.* Comparative chromosome painting of chicken autosomal paints 1–9 in nine different bird species. *Cytogenet. Genome Res.* **103**, 173–184 (2003).

136. Derjusheva, S., Kurganova, A., Habermann, F. & Gaginskaya, E. High chromosome conservation detected by comparative chromosome painting in chicken, pigeon and passerine birds. *Chromosome Res.* **12**, 715–723 (2004).

137. Jaari, S., Li, M.-H. & Merilä, J. A first-generation microsatellite-based genetic linkage map of the Siberian jay (Perisoreus infaustus): insights into avian genome evolution. *BMC Genomics* vol. 10 (2009).

138. Schlötterer, C. The evolution of molecular markers--just a matter of fashion? *Nat. Rev. Genet.* **5**, 63–69 (2004).

139. Wink, M. Use of DNA markers to study bird migration. *J. Ornithol.* **147**, 234–244 (2006).

140. Backstrom, N., Karaiskou, N., Leder, E. H. & Gustafsson, L. A Gene-Based Genetic Linkage Map of the Collared Flycatcher (Ficedula albicollis) Reveals Extensive Synteny and Gene-Order Conservation During 100 Million …. *Genetics* (2008).

141. Stapley, J., Birkhead, T. R., Burke, T. & Slate, J. A linkage map of the zebra finch *Taeniopygia guttata* provides new insights into avian genome evolution. *Genetics* **179**, 651–667 (2008).

142. Spinelli, J. B. & Haigis, M. C. The multifaceted contributions of mitochondria to cellular metabolism. *Nat. Cell Biol.* **20**, 745–754 (2018).

143. Vyas, S., Zaganjor, E. & Haigis, M. C. Mitochondria and Cancer. *Cell* **166**, 555–566 (2016).

144. Sagan, L. On the origin of mitosing cells. *J. Theor. Biol.* **14**, 255–274 (1967).

145. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).

146. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

147. Sengupta, S., Yang, X. & Higgs, P. G. The mechanisms of codon reassignments in mitochondrial genetic codes. *J. Mol. Evol.* **64**, 662–688 (2007).

148. Smith, D. R. & Keeling, P. J. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10177–10184 (2015).

149. Kolesnikov, A. A. & Gerasimov, E. S. Diversity of mitochondrial genome organization. *Biochemistry* **77**, 1424–1435 (2012).

150. McKinney, E. A. & Oliveira, M. T. Replicating animal mitochondrial DNA. *Genet. Mol. Biol.* **36**, 308–315 (2013).

151. Gibb, G. C., Kardailsky, O., Kimball, R. T., Braun, E. L. & Penny, D. Mitochondrial genomes and avian phylogeny: complex characters and resolvability without explosive radiations. *Mol. Biol. Evol.* **24**, 269–280 (2007).

152. Satoh, T. P., Miya, M., Mabuchi, K. & Nishida, M. Structure and variation of the mitochondrial genome of fishes. *BMC Genomics* **17**, 719 (2016).

153. Mackiewicz, P., Urantówka, A. D., Kroczak, A. & Mackiewicz, D. Resolving Phylogenetic Relationships within Passeriformes Based on Mitochondrial Genes and Inferring the Evolution of Their Mitogenomes in Terms of Duplications. *Genome Biol. Evol.* **11**, 2824–2849 (2019).

154. Heyer, E. *et al.* Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am. J. Hum. Genet.* **69**, 1113–1126 (2001).

155. Bronstein, O., Kroh, A. & Haring, E. Mind the gap! The mitochondrial control region and its power as a phylogenetic marker in echinoids. *BMC Evol. Biol.* **18**, 80 (2018).

156. DeSalle, R. & Hadrys, H. Evolutionary biology and mitochondrial genomics: 50 000 mitochondrial DNA genomes and counting. *eLS* 1–35 (2017) doi:10.1002/9780470015902.a0027270.

157. Formenti, G. *et al.* Complete vertebrate mitogenomes reveal widespread gene duplications and repeats. *Cold Spring Harbor Laboratory* 2020.06.30.177956 (2020) doi:10.1101/2020.06.30.177956.

158. Tamashiro, R. A. *et al.* What are the roles of taxon sampling and model fit in tests of cyto-nuclear discordance using avian mitogenomic data? *Mol. Phylogenet. Evol.* **130**, 132–142 (2019).

159. Formenti, G. *et al.* Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biol.* **22**, 120 (2021).

160. Allio, R., Donega, S. & Galtier, N. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Mol. Biol.* **34**, 2762-2772 (2017).

161. Gissi, C., Iannelli, F. & Pesole, G. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity* **101**, 301–320 (2008).

162. Irestedt, M. & Ohlson, J. I. The division of the major songbird radiation into Passerida and 'core Corvoidea' (Aves: Passeriformes) — the species tree vs. gene trees. *Zool. Scr.* **37**, 305–313 (2008).

163. Moore, W. S. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution* **49**, 718–726 (1995).

164. Gibb, G. C. *et al.* New Zealand Passerines Help Clarify the Diversification of Major Songbird Lineages during the Oligocene. *Genome Biol. Evol.* **7**, 2983–2995 (2015).

165. Mindell, D. P., Fuchs, J. & Johnson, J. A. Phylogeny, Taxonomy, and Geographic Diversity of Diurnal Raptors: Falconiformes, Accipitriformes, and Cathartiformes. in *Birds of Prey: Biology and conservation in the XXI century* (eds. Sarasola, J. H., Grande, J. M. & Negro, J. J.) 3–32 (Springer International Publishing, 2018). doi:10.1007/978-3-319-73745-4_1.

166. Hong, J. H. *et al.* Ancient mitochondrial DNA analysis of avian bones collected from the 4th century pit burial found in South Korea. *Archaeological Research in Asia* **24**, 100214 (2020).

167. Pacheco, M. A. *et al.* Mode and Rate of Evolution of Haemosporidian Mitochondrial Genomes: Timing the Radiation of Avian Parasites. *Mol. Biol. Evol.* **35**, 383–403 (2018).

168. Rutkowski, R. *et al.* Conservation genetics of the CapercaillieTetrao urogallusin Poland — diversity of mitochondrial DNA in remnant and extinct populations. *Acta Ornithol.* **52**, 179–196 (2017).

169. Dalloul, R. A. *et al.* Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol.* **8**, (2010).

170. Huang, Y. *et al.* The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat. Genet.* **45**, 776–783 (2013).

171. Warren, W. C. *et al.* The genome of a songbird. *Nature* **464**, 757–762 (2010).

172. Balakrishnan, C. N., Edwards, S. V. & Clayton, D. F. The Zebra Finch genome and avian genomics in the wild. *Emu - Austral Ornithology* **110**, 233–241 (2010).

173. Zhang, G. *et al.* Comparative genomic data of the Avian Phylogenomics Project. *Gigascience* **3**, 26 (2014).

174. Shapiro, M. D. *et al.* Genomic Diversity and Evolution of the Head Crest in the Rock Pigeon. *Science* vol. 339 1063–1067 (2013).

175. Zhang, G., Jarvis, E. D. & Gilbert, M. T. P. A flock of genomes. *Science* **346**, 1308–1309 (2014).

176. Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).

177. Mayr, E. Peters's 'Check-list of Birds of the World'. *Auk* **54**, 550–551 (1937).

178. Livezey, B. C. & Zusi, R. L. Higher-order phylogeny of modern birds (Theropoda, Aves: Neornithes) based on comparative anatomy. II. Analysis and discussion. *Zool. J. Linn. Soc.* **149**, 1–95 (2007).

179. SIBLEY & CG. Phylogeny and Classification of Birds. *A Study in Molecular Evolution* (1990).

180. Pratt, R. C. *et al.* Toward resolving deep neoaves phylogeny: data, signal enhancement, and priors. *Mol. Biol. Evol.* **26**, 313–326 (2009).

181. Pacheco, M. A. *et al.* Evolution of modern birds revealed by mitogenomics: timing the radiation and origin of major orders. *Mol. Biol. Evol.* **28**, 1927–1942 (2011).

182. Ericson, P. G. P. *et al.* Diversification of Neoaves: integration of molecular sequence data and fossils. *Biol. Lett.* **2**, 543–547 (2006).

183. Hackett, S. J. *et al.* A Phylogenomic Study of Birds Reveals Their Evolutionary History. *Science* vol. 320 1763–1768 (2008).

184. Wang, N., Braun, E. L. & Kimball, R. T. Testing hypotheses about the sister group of the passeriformes using an independent 30-locus data set. *Mol. Biol. Evol.* **29**, 737–750 (2012).

185. Kimball, R. T., Wang, N., Heimer-McGinn, V., Ferguson, C. & Braun, E. L. Identifying localized biases in large datasets: a case study using the avian tree of life. *Mol. Phylogenet. Evol.* **69**, 1021–1032 (2013).

186. McCormack, J. E. *et al.* A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS One* **8**, e54848 (2013).

187. Suh, A. *et al.* Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nat. Commun.* **2**, 443 (2011).

188. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**, 225–231 (2006).

189. Song, S., Liu, L., Edwards, S. V. & Wu, S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 14942–14947 (2012).

190. Chojnowski, J. L., Kimball, R. T. & Braun, E. L. Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes. *Gene* **410**, 89–96 (2008).

191. Patel, S., Kimball, R. T. & Braun, E. L. Error in phylogenetic estimation for bushes in the tree of life. *J. Phylogenet. Evol. Biol* **1**, 2 (2013).

192. Mayr, G. Metaves, Mirandornithes, Strisores and other novelties - a critical review of the higher-level phylogeny of neornithine birds. *J. Zoolog. Syst. Evol. Res.* **49**, 58–76 (2011).

193. Nabholz, B., Künstner, A., Wang, R., Jarvis, E. D. & Ellegren, H. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol. Biol. Evol.* **28**, 2197–2210 (2011).

194. Matzke, A. *et al.* Retroposon insertion patterns of neoavian birds: strong evidence for an extensive incomplete lineage sorting era. *Mol. Biol. Evol.* **29**, 1497–1501 (2012).

195. Wolf, Y. I., Rogozin, I. B., Grishin, N. V. & Koonin, E. V. Genome trees and the tree of life. *Trends Genet.* **18**, 472–479 (2002).

196. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804 (2003).

197. Twyford, A. D. & Ennos, R. A. Next-generation hybridization and introgression. *Heredity* **108**, 179–189 (2012).

198. Benton, M. J. The origins of modern biodiversity on land. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 3667–3679 (2010).

199. Schulte, P. *et al.* The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. *Science* **327**, 1214–1218 (2010).

200. Clancey, E., Des Roches, S. & Eastman, J. M. Ecological opportunity and the origin of adaptive radiations. *Journal of* (2010).

201. Longrich, N. R., Tokaryk, T. & Field, D. J. Mass extinction of birds at the Cretaceous-Paleogene (K-Pg) boundary. *Proceedings of the National Academy of Sciences* vol. 108 15253–15257 (2011).

202. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

203. Meredith, R. W., Zhang, G., Gilbert, M. T. P., Jarvis, E. D. & Springer, M. S. Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science* **346**, 1254390 (2014).

204. Opazo, J. C. *et al.* Gene Turnover in the Avian Globin Gene Families and Evolutionary Changes in Hemoglobin Isoform Expression. *Molecular Biology and Evolution* vol. **32**, 871–887 (2015).

205. Ribeiro, Â. M. *et al.* A refined model of the genomic basis for phenotypic variation in vertebrate hemostasis. *BMC Evol. Biol.* **15**, 124 (2015).

206. Pfenning, A. R. *et al.* Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* (2014) doi:10.1126/science.1256846.

207. Whitney, O. *et al.* Core and region-enriched networks of behaviorally regulated genes and the singing genome. *Science* **346**, 1256780 (2014).

208. Wirthlin, M., Lovell, P. V., Jarvis, E. D. & Mello, C. V. Comparative genomics reveals molecular features unique to the songbird lineage. *BMC Genomics* **15**, 1082 (2014).

209. Korlach, J. *et al.* 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 (2017).

210. Farre, M., Narayan, J., Slavov, G. T. & Damas, J. Novel insights into chromosome evolution in birds, archosaurs, and reptiles. *Genome Biol.* **8**, 2442-2451 (2016).

211. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).

212. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).

213. Feng, S. *et al.* Author Correction: Dense sampling of bird diversity increases power of comparative genomics. *Nature* **592**, E24 (2021).

214. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4325–4333 (2018).

215. Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences* **119**, e2115642118 (2022).

216. Blaxter, M. *et al.* Why sequence all eukaryotes? *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).

217. Teeling, E. C. *et al.* Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. *Annu Rev Anim Biosci* **6**, 23–46 (2018).

218. Formenti, G. *et al.* The era of reference genomes in conservation genomics. *Trends Ecol. Evol.* (2022) doi:10.1016/j.tree.2021.11.008.

219. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).

220. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

221. Kronenberg, Z. N. *et al.* Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat. Commun.* **12**, 1935 (2021).

222. Howe, K. *et al.* Significantly improving the quality of genome assemblies through curation. *Gigascience* **10**, (2021).

223. Jebb, D. *et al.* Six reference-quality genomes reveal evolution of bat adaptations. *Nature* **583**, 578–584 (2020).

224. Morin, P. A. *et al.* Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Mol. Ecol. Resour.* **21**, 1008–1020 (2021).

225. Theofanopoulou, C., Gedman, G., Cahill, J. A., Boeckx, C. & Jarvis, E. D. Universal nomenclature for oxytocin-vasotocin ligand and receptor families. *Nature* **592**, 747–755 (2021).

226. Yang, C. *et al.* Evolutionary and biomedical insights from a marmoset diploid genome assembly. *Nature* **594**, 227–233 (2021).

227. Zhou, Y. *et al.* Platypus and echidna genomes reveal mammalian biology and evolution. *Nature* **592**, 756–762 (2021).

228. Chow, W. *et al.* gEVAL — a web-based browser for evaluating genome assemblies. *Bioinformatics* **32**, 2508–2510 (2016).

229. Stewart, J. R., Lister, A. M., Barnes, I. & Dalén, L. Refugia revisited: individualistic responses of species in space and time. *Proc. Biol. Sci.* **277**, 661–671 (2010).

230. Montgomery, W. I., Provan, J., McCabe, A. M. & Yalden, D. W. Origin of British and Irish mammals: disparate post-glacial colonisation and species introductions. *Quat. Sci. Rev.* **98**, 144–165 (2014).

231. Miga, K. H. & Wang, T. The Need for a Human Pangenome Reference Sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).

232. Brandt, D. Y. C. *et al.* Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3* **5**, 931–941 (2015).

233. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).

234. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* **15**, e1008302 (2019).

235. Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome Biol.* **20**, 159 (2019).

236. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).

237. International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).

238. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

239. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).

240. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).

241. Martiniano, R., Garrison, E., Jones, E. R., Manica, A. & Durbin, R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol.* **21**, 250 (2020).

242. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).

243. Zhou, B. *et al.* AntCaller: an accurate variant caller incorporating ancient DNA damage. *Mol. Genet. Genomics* **292**, 1419–1430 (2017).

244. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).

245. The Computational Pan-Genomics Consortium *et al.* Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* **19**, 118–135 (2016).

246. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13950–13955 (2005).

247. Francis, W. R. & Wörheide, G. Similar Ratios of Introns to Intergenic Sequence across Animal Genomes. *Genome Biol. Evol.* **9**, 1582–1598 (2017).

248. Gerdol, M. *et al.* Massive gene presence/absence variation in the mussel genome as an adaptive strategy: first evidence of a pan-genome in Metazoa. *bioRxiv* 781377 (2019) doi:10.1101/781377.

249. Tian, X. *et al.* Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Sci. China Life Sci.* **63**, 750–763 (2020).

250. Heaton, M. P. *et al.* A Reference Genome Assembly of Simmental Cattle, *Bos taurus taurus. J. Hered.* **112**, 184–191 (2021).

251. Wang, K. *et al.* The Chicken Pan-Genome Reveals Gene Content Variation and a Promoter Region Deletion in IGF2BP1 Affecting Body Size. *Mol. Biol. Evol.* **38**, 5066–5081 (2021).

252. Li, M. *et al.* De novo assembly of 20 chicken genomes reveals the undetectable phenomenon for thousands of core genes on micro-chromosomes and sub-telomeric regions. *Mol. Biol. Evol.* (2022) doi:10.1093/molbev/msac066.

253. Rakocevic, G. *et al.* Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**, 354–362 (2019).

254. Jain, C., Misra, S., Zhang, H., Dilthey, A. & Aluru, S. Accelerating Sequence Alignment to Graphs. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 451–461 (2019). doi:10.1109/IPDPS.2019.00055.

255. Rautiainen, M., Mäkinen, V. & Marschall, T. Bit-parallel sequence-to-graph alignment. *Bioinformatics* **35**, 3599–3607 (2019).

256. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).

257. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *bioRxiv* 2021.11.10.467921 (2021) doi:10.1101/2021.11.10.467921.

258. Beyer, W. *et al.* Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics* **35**, 5318–5320 (2019).

259. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).

260. Barnosky, A. D. *et al.* Has the Earth's sixth mass extinction already arrived? *Nature* vol. 471 51–57 (2011).

261. Supple, M. A. & Shapiro, B. Conservation of biodiversity in the genomics era. *Genome Biol.* **19**, 131 (2018).

262. Turner, A. K. *The Barn Swallow*. (T. & A. D. Poyser, 2006).

263. Møller, A. P. Sexual selection and the barn swallow.- Oxford Univ. (1994).

264. Turner, A. & Rose, C. *A Handbook to the Swallows and Martins of the World*. (A&C Black, 2010).

265. Moller, A. P., de Lope, F. & Saino, N. Sexual selection in the barn swallow *Hirundo rustica*. VI. Aerodynamic adaptations. *J. Evol. Biol.* **8**, 671–687 (1995).

266. Saino, N., Romano, M., Ambrosini, R., Ferrari, R. P. & Møller, A. P. Timing of Reproduction and Egg Quality Covary with Temperature in the insectivorous Barn Swallow, *Hirundo rustica*. *Funct. Ecol.* **18**, 50–57 (2004).

267. Snow, D. W. & Cramp, S. *The Complete Birds of the Western Palearctic*. (Oxford University Press, 1998).

268. Scandolara, C. *et al. Context-, phenotype-, and kin-dependent natal dispersal of barn swallows (Hirundo rustica*). *Behav. Ecol.* **25**, 180–190 (2013).

269. Shirihai, H., Dovrat, E., Christie, D. A. & Harris, A. *The birds of Israel*. vol. 876 (Academic Press London, 1996).

270. Turner, A. & Rose, C. *Swallows and martins :an identification guide and handbook /*. http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=FCL.xis&method=post&formato=2&cantidad=1&expresion=mfn=000459 (1989).

271. Saino, N. *et al.* Timing of molt of barn swallows is delayed in a rare Clock genotype. *PeerJ* vol. 1 e17 (2013).

272. Zink, R. M., Pavlova, A., Rohwer, S. & Drovetski, S. V. Barn swallows before barns: population histories and intercontinental colonization. *Proc. Biol. Sci.* **273**, 1245–1251 (2006).

273. Baker, P. J. & Harris, S. Urban mammals: what does the future hold? An analysis of the factors affecting patterns of use of residential gardens in Great Britain. *Mamm. Rev.* **37**,297-315 (2007).

274. Hulme-Beaman, A., Dobney, K., Cucchi, T. & Searle, J. B. An Ecological and Evolutionary Framework for Commensalism in Anthropogenic Environments. *Trends Ecol. Evol.* **31**, 633–645 (2016).

275. Scordato, E. S. C. & Safran, R. J. Geographic variation in sexual selection and implications for speciation in the Barn Swallow. *Avian Research* **5**, 8 (2014).

276. Smith, C. C. R., Flaxman, S. M. & Scordato, E. S. C. Demographic inference in barn swallows using whole-genome data shows signal for bottleneck and subspecies differentiation during the Holocene. *Molecular* (2018).

277. Ambrosini, R. *et al.* Maintenance of livestock farming may buffer population decline of the Barn Swallow *Hirundo rustica*. *Bird Conserv. Int.* **22**, 411–428 (2012).

278. Møller, A. P. Parallel declines in abundance of insects and insectivorous birds in Denmark over 22 years. *Ecol. Evol.* **9**, 6581–6587 (2019).

279. Darwin, C. & Murray, J. The Variation of Animals and Plants Under Domestication edit. (1868).

280. Belyaev, D. K. & Khvostova, V. V. Domestication, plant and animal. *Encyclopaedia Britannica* **15**, 936–942 (1974).

281. Wilkins, A. S., Wrangham, R. W. & Fitch, W. T. The 'Domestication Syndrome' in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics. *Genetics* **197**, 795–808 (2014).

282. O'Rourke, T. & Boeckx, C. Glutamate receptors in domestication and modern human evolution. *Neurosci. Biobehav. Rev.* **108**, 341–357 (2020).

283. Herman, J. P., Mueller, N. K. & Figueiredo, H. Role of GABA and glutamate circuitry in hypothalamo-pituitary-adrenocortical stress integration. *Ann. N. Y. Acad. Sci.* **1018**, 35–45 (2004).

284. O'Rourke, T. *et al.* Capturing the Effects of Domestication on Vocal Learning Complexity. *Trends Cogn. Sci.* **25**, 462–474 (2021).

285. Shi, Z. *et al.* miR-9 regulates basal ganglia-dependent developmental vocal learning and adult vocal performance in songbirds. *Elife* **7**, (2018).

286. Teramitsu, I., Kudo, L. C., London, S. E., Geschwind, D. H. & White, S. A. Parallel FoxP1 and FoxP2 expression in songbird and human brain predicts functional interaction. *J. Neurosci.* **24**, 3152–3163 (2004).

287. Teramitsu, I. & White, S. A. FoxP2 regulation during undirected singing in adult songbirds. *J. Neurosci.* **26**, 7390–7394 (2006).

288. Teramitsu, I., Poopatanapong, A., Torrisi, S. & White, S. A. Striatal FoxP2 is actively regulated during songbird sensorimotor learning. *PLoS One* **5**, e8548 (2010).

289. Pap, P. L. *et al.* Sexual Dimorphism and Population Differences in Structural Properties of Barn Swallow (Hirundo rustica) Wing and Tail Feathers. *PLoS One* **10**, e0130844 (2015).

290. Pap, P. L. *et al.* Selection on multiple sexual signals in two Central and Eastern European populations of the barn swallow. *Ecol. Evol.* **9**, 11277–11287 (2019).

291. Saino, N. *et al.* A trade-off between reproduction and feather growth in the barn swallow (*Hirundo rustica*). *PLoS One* **9**, e96428 (2014).

292. Saino, N. *et al.* Migration phenology and breeding success are predicted by methylation of a photoperiodic gene in the barn swallow. *Sci. Rep.* **7**, 45412 (2017).

293. Saino, N. *et al.* Wing morphology, winter ecology, and fecundity selection: evidence for sex-dependence in barn swallows (*Hirundo rustica*). *Oecologia* **184**, 799–812 (2017).

294. The Barn Swallow. (2010) doi:10.5040/9781472596888.

295. Gwinner, E. Circannual rhythms in birds. *Curr. Opin. Neurobiol.* **13**, 770–778 (2003).

296. Wikelski, M. *et al.* Avian circannual clocks: adaptive significance and possible involvement of energy turnover in their proximate control. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 411–423 (2008).

297. Visser, M. E., Caro, S. P., van Oers, K., Schaper, S. V. & Helm, B. Phenology, seasonal timing and circannual rhythms: towards a unified framework. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 3113–3127 (2010).

298. Merrow, M., Spoelstra, K. & Roenneberg, T. The circadian cycle: daily rhythms from behaviour to genes. *EMBO Rep.* **6**, 930–935 (2005).

299. Liedvogel, M. & Lundberg, M. The genetics of animal movement and migration syndromes. in *Animal movement across scales* 219–231 (Oxford University Press, 2014).

300. Pulido, F. & Berthold, P. Quantitative Genetic Analysis of Migratory Behaviour. in *Avian Migration* 53–77 (Springer Berlin Heidelberg, 2003). doi:10.1007/978-3-662-05957-9_4.

301. Liedvogel, M., Szulkin, M., Knowles, S. C. L., Wood, M. J. & Sheldon, B. C. Phenotypic correlates of Clock gene variation in a wild blue tit population: evidence for a role in seasonal timing of reproduction. *Mol. Ecol.* **18**, 2444–2456 (2009).

302. Caprioli, M. *et al.* Clock gene variation is associated with breeding phenology and maybe under directional selection in the migratory barn swallow. *PLoS One* **7**, e35140 (2012).

303. Saino, N. *et al.* Polymorphism at the Clock gene predicts phenology of long-distance migration in birds. *Mol. Ecol.* **24**, 1758–1773 (2015).

304. Bazzi, G. *et al.* Adcyap1 polymorphism covaries with breeding latitude in a Nearctic migratory songbird, the Wilson's warbler (*Cardellina pusilla*). *Ecology and Evolution* vol. 6 3226–3239 (2016).

305. O'Malley, K. G., Ford, M. J. & Hard, J. J. Clock polymorphism in Pacific salmon: evidence for variable selection along a latitudinal gradient. *Proc. Biol. Sci.* **277**, 3703–3714 (2010).

306. Chakarov, N., Jonker, R. M., Boerner, M., Hoffman, J. I. & Krüger, O. Variation at phenological candidate genes correlates with timing of dispersal and plumage morph in a sedentary bird of prey. *Mol. Ecol.* **22**, 5430–5440 (2013).

307. Bourret, A. & Garant, D. Candidate gene-environment interactions and their relationships with timing of breeding in a wild bird population. *Ecol. Evol.* **5**, 3628–3641 (2015).

308. Safran, R. J. *et al.* Genome-wide differentiation in closely related populations: the roles of selection and geographic isolation. *Mol. Ecol.* **25**, 3865–3883 (2016).

309. Primmer, C. R., Møller, A. P. & Ellegren, H. Resolving genetic relationships with microsatellite markers: a parentage testing system for the swallow Hirundo rustica. *Mol. Ecol.* **4**, 493–498 (1995).

310. Tsyusko, O. V. *et al.* Microsatellite markers isolated from barn swallows (*Hirundo rustica*). *Mol. Ecol. Notes* **7**, 833–835 (2007).

311. von Rönn, J. A. C., Shafer, A. B. A. & Wolf, J. B. W. Disruptive selection without genome-wide evolution across a migratory divide. *Mol. Ecol.* **25**, 2529–2541 (2016).

312. Safran, R. J. *et al.* The maintenance of phenotypic divergence through sexual selection: An experimental study in barn swallows *Hirundo rustica*. *Evolution* **70**, 2074–2084 (2016).

313. Scordato, E. S. C. *et al.* Migratory divides coincide with reproductive barriers across replicated avian hybrid zones above the Tibetan Plateau. *Ecol. Lett.* **23**, 231–241 (2020).

314. Turbek, S. P. *et al.* A migratory divide spanning two continents is associated with genomic and ecological divergence. *Evolution* (2022) doi:10.1111/evo.14448.

315. Scordato, E. S. C. *et al.* Genomic variation across two barn swallow hybrid zones reveals traits associated with divergence in sympatry and allopatry. *Mol. Ecol.* **26**, 5676–5691 (2017).

316. Smith, C. C. R. *et al.* Demographic inference in barn swallows using whole-genome data shows signal for bottleneck and subspecies differentiation during the Holocene. *Mol. Ecol.* **27**, 4200–4212 (2018).

317. Schield, D. R. *et al.* Sex-linked genetic diversity and differentiation in a globally distributed avian species complex. *Mol. Ecol.* **30**, 2313–2332 (2021).

318. Formenti, G. *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* **8**, (2019).

319. 10k, B. *et al.* Home. https://b10k.genomics.cn/.

320. Feng, S. *et al.* Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252–257 (2020).

321. Sheldon, F. H., Whittingham, L. A., Moyle, R. G., Slikas, B. & Winkler, D. W. Phylogeny of swallows (Aves: Hirundinidae) estimated from nuclear and mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* **35**, 254–270 (2005).

322. Dor, R., Safran, R. J., Sheldon, F. H., Winkler, D. W. & Lovette, I. J. Phylogeny of the genus *Hirundo* and the Barn Swallow subspecies complex. *Mol. Phylogenet. Evol.* **56**, 409–418 (2010).

323. Malaitad, T., Laipasu, P., Eiamampai, K., Poeaim, S. & Others. Identification of the Subspecies and Gender of Barn Swallow (*Hirundo rustica*). *International Journal of Agricultural Technology* **12**, 1549–1556 (2016).

324. Carter, J. K. *et al.* Complete mitochondrial genomes provide current refined phylogenomic hypotheses for relationships among ten Hirundo species. *Mitochondrial DNA B Resour* **5**, 2881–2885 (2020).

325. Hagemeijer, W. J. M. & Blair, M. J. The EBCC atlas of European breeding birds. *Poyser, London* **479**, (1997).

326. Cleere, N. *Nightjars, Potoos, Frogmouths, Oilbird, and Owlet-nightjars of the World*. (Princeton University Press, 2021). doi:10.1515/9781400836161.

327. Cramp, S. & Brooks, D. J. Vol. IV: Terns to woodpeckers. (1985).

328. Cleere, N. *Nightjars: A Guide to Nightjars and related birds*. (A&C Black, 2010).

329. Woods, C. P., Czenze, Z. J. & Brigham, R. M. The avian 'hibernation' enigma: thermoregulatory patterns and roost choice of the common poorwill. *Oecologia* **189**, 47–53 (2019).

330. Carey, C. *Life In The Cold: Ecological, Physiological, And Molecular Mechanisms*. (CRC Press, 2019).

331. French, A. R. Hibernation in birds: comparisons with mammals. in *Life in the Cold* 43–53 (CRC Press, 2019).

332. Brinkløv, S., Fenton, M. B. & Ratcliffe, J. M. Echolocation in Oilbirds and swiftlets. *Front. Physiol.* **4**, 123 (2013).

333. Cresswell, B. & Edwards, D. Geolocators reveal wintering areas of European Nightjar (Caprimulgus europaeus). *Bird Study* **60**, 77–86 (2013).

334. Norevik, G., Åkesson, S. & Hedenström, A. Migration strategies and annual space-use in an Afro-Palaearctic aerial insectivore - the European nightjar *Caprimulgus europaeus*. *J. Avian Biol.* **48**, 738–747 (2017).

335. Langston, R. H. W. *et al.* Nightjar *Caprimulgus europaeus* and Woodlark *Lullula arborea*- recovering species in Britain? *Ibis* **149**, 250–260 (2007).

336. Eaton, M. *et al.* Birds of Conservation Concern 4: the population status of birds in the UK, Channel Islands and Isle of Man. *Br. Birds* **108**, 708–746 (2015).

337. Keller, V., Gerber, A., Schmid, H., Volet, B. & Zbinden, N. Rote Liste Brutvögel. Gefährdete Arten der Schweiz, Stand 2010. Umwelt-Vollzug Nr. 1019. *Bundesamt für Umwelt, Bern, und Schweizerische Vogelwarte, Sempach* (2010).

338. Winiger, N., Korner, P., Arlettaz, R. & Jacot, A. Vegetation structure and decreased moth abundance limit the recolonisation of restored habitat by the European Nightjar. *Rethinking ecology* **3**, 15 (2018).

339. Evens, R., Beenaerts, N., Witters, N. & Artois, T. Study on the foraging behaviour of the European nightjar *Caprimulgus europaeus* reveals the need for a change in conservation strategy in Belgium. *J. Avian Biol.* **48**, 1238–1245 (2017).

340. BirdLife International. Species factsheet: *Caprimulgus europaeus*. Downloaded from http://www.birdlife.org on 07/02/2022. (2022).

341. Sierro, A. & Erhardt, A. Light pollution hampers recolonization of revitalised European Nightjar habitats in the Valais (Swiss Alps). *J. Ornithol.* **160**, 749–761 (2019).

342. Langston, R. H. W., Liley, D., Murison, G., Woodfield, E. & Clarke, R. T. What effects do walkers and dogs have on the distribution and productivity of breeding European Nightjar *Caprimulgus europaeus*? *Ibis* **149**, 27–36 (2007).

343. Lowe, A., Rogers, A. C. & Durrant, K. L. Effect of human disturbance on long-term habitat use and breeding success of the European Nightjar, *Caprimulgus europaeus*. *Avian Conservation and Ecology* vol. 9 (2014).

344. Louette, G. *et al.* Bridging the gap between the Natura 2000 regional conservation status and local conservation objectives. *J. Nat. Conserv.* **19**, 224–235 (2011).

345. Polakowski, M., Broniszewska, M., Kirczuk, L. & Kasprzykowski, Z. Habitat Selection by the European Nightjar Caprimulgus europaeus in North-Eastern Poland: Implications for Forest Management. *For. Trees Livelihoods* **11**, 291 (2020).

346. Larsen, C., Speed, M., Harvey, N. & Noyes, H. A. A molecular phylogeny of the nightjars (Aves: Caprimulgidae) suggests extensive conservation of primitive morphological traits across multiple lineages. *Mol. Phylogenet. Evol.* **42**, 789–796 (2007).

347. Han, K.-L., Robbins, M. B. & Braun, M. J. A multi-gene estimate of phylogeny in the nightjars and nighthawks (Caprimulgidae). *Mol. Phylogenet. Evol.* **55**, 443–453 (2010).

348. Lawrie, Y., Swann, R., Stronach, P., Perlman, Y. & Collinson, J. M. The taxonomic position and breeding range of Golden Nightjar Caprimulgus eximius (Caprimulgidae). *Ostrich* **88**, 281–286 (2017).

349. Schweizer, M. *et al.* A molecular analysis of the mysterious Vaurie's Nightjar *Caprimulgus centralasicus* yields fresh insight into its taxonomic status. *Journal of Ornithology* vol. 161 635–650 (2020).

350. Brown, L. & Amadon, D. *Eagles, Hawks, and Falcons of the World*. (McGraw-Hill, 1968).

351. Cramp, S. & Simmons, R. The birds of Western Palearctic, the Middle East and North Africa, vol II. (1980).

352. Newton, I. *Population ecology of raptors*. (A&C Black, 2010).

353. Cramp, S. & Simmons, K. E. L. Handbook of the Birds of Europe, the Middle East and North Africa. The Birds of the Western Palearctic. Vol. II. Hawks to Bustards.- Oxford University Press. (1980).

354. Negro, J. J., De la Riva, M. & Bustamante, J. Patterns of winter distribution and abundance of sedentary lesser Kestrel (*Falco naumanni*) in Spain. (1991).

355. Sarà, M. *et al. Broad-front migration leads to strong migratory connectivity in the lesser kestrel (Falco naumanni*). *J. Biogeogr*. **46**, 2663–2677 (2019).

356. Cramp, S. *The complete birds of the Western Palearctic on CD-ROM*. (Oxford University Press, 1998).

357. BirdLife International. *Birds in Europe: Population Estimates, Trends and Conservation Status*. (BirdLife International, 2004).

358. Donald, P. F., Green, R. E. & Heath, M. F. Agricultural intensification and the collapse of Europe's farmland bird populations. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**, 25–29 (2001).

359. Rosenberg, K. V. *et al.* Decline of the North American avifauna. *Science* **366**, 120–124 (2019).

360. Iñigo, A. & Barov, B. Species Action Plan for the Lesser Kestrel *Falco naumanni* in the European Union. SEO/BirdLife and BirdLife International for the European Commission. (2011).

361. BirdLife International. *European birds of conservation concern: populations, trends and national responsibilities*. (BirdLife International, Cambridge, UK, 2017).

362. Cecere, J. G. *et al.* Spatial segregation of home ranges between neighbouring colonies in a diurnal raptor. *Scientific Reports* vol. 8 (2018).

363. Podofillini, S. *et al.* Home, dirty home: effect of old nest material on nest-site selection and breeding performance in a cavity-nesting raptor. *Curr. Zool.* **64**, 693–702 (2018).

364. Podofillini, S. *et al.* Benefits of extra food to reproduction depend on maternal condition. *Oikos* **128**, 943–959 (2019).

365. Podofillini, S. Short and long-term effects of variation in the breeding environment on behaviour and fitness traits in acolonial, cavity nesting raptor. PhD Thesis (2019).

366. Cecere, J. G. *et al.* Inter-individual differences in foraging tactics of a colonial raptor: consistency, weather effects, and fitness correlates. *Mov Ecol* **8**, 28 (2020).

367. Romano, A., Corti, M., Soravia, C., Cecere, J. G. & Rubolini, D. Ectoparasites exposure affects early growth and mouth colour in nestlings of a cavity-nesting raptor. *Behav. Ecol. Sociobiol.* **75**, (2021).

368. Assandri, G. *et al.* Context-dependent foraging habitat selection in a farmland raptor along an agricultural intensification gradient. *Agric. Ecosyst. Environ.* **326**, 107782 (2022).

369. Ward, D., Pepler, D. & Botha, R. The influence of sample size on the determination of population trends in the vulnerable Lesser Kestrel *Falco naumanni* overwintering in South Africa. *Ostrich* **79**, 199–204 (2008).

370. Gustin, M. *et al.* Conservation of the Lesser Kestrel *Falco Naumanni* in and around the Alta Murgia National Park: scientific results allow for tailored plans. in *Acts of the Workshop of the Alta Murgia National Park, Gravina in Puglia* (2017).

371. Bounas, A. *et al.* Genetic structure of a patchily-distributed philopatric migrant: implications for management and conservation. *bioRxiv* 216069 (2017) doi:10.1101/216069.

372. Di Maggio, R., Campobello, D. & Sarà, M. Lesser kestrel diet and agricultural intensification in the Mediterranean: An unexpected win-win solution? *J. Nat. Conserv.* **45**, 122–130 (2018).

373. Morganti, M., Ambrosini, R. & Sarà, M. Different trends of neighboring populations of Lesser Kestrel: Effects of climate and other environmental conditions. *Popul. Ecol.* **61**, 300–314 (2019).

374. Campobello, D., Di Maggio, R. & Sara, M. Planning conservation actions by investigating nest preferences and biotic and abiotic factors within lesser kestrel (*Falco naumanni*) colonies. in *Linking behaviour to populations and communities: how can behavioural ecology inform conservation?* (2019).

375. Morganti, M. *et al.* Assessing the relative importance of managed crops and semi-natural grasslands as foraging habitats for breeding lesser kestrels *Falco naumanni* in southeastern Italy. *wbio* **2021**, (2021).

376. Ortego, J., Calabuig, G., Aparicio, J. M. & Cordero, P. J. Genetic consequences of natal dispersal in the colonial lesser kestrel. *Mol. Ecol.* **17**, 2051–2059 (2008).

377. Alcaide, M., Edwards, S. V., Negro, J. J., Serrano, D. & Tella, J. L. Extensive polymorphism and geographical variation at a positively selected MHC class II B gene of the lesser kestrel (*Falco naumanni*). *Mol. Ecol.* **17**, 2652–2665 (2008).

378. Di Maggio, R. *et al.* Do not disturb the family: roles of colony size and human disturbance in the genetic structure of lesser kestrel. *J. Zool.* **295**, 108–115 (2015).

379. Corti, M. *et al.* Sequence variation in melanocortin-1-receptor and tyrosinase-related protein 1 genes and their relationship with melanin-based plumage trait expression in Lesser Kestrel (*Falco naumanni*) males. *J. Ornithol.* **159**, 587–591 (2018).

380. Bounas, A. *et al.* Using genetic markers to unravel the origin of birds converging towards pre-migratory sites. *Sci. Rep.* **8**, 8326 (2018).

381. Schiffels, S. & Wang, K. MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. *Methods Mol. Biol.* **2090**, 147–166 (2020).

382. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 2020.05.22.110833 (2020) doi:10.1101/2020.05.22.110833.

383. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).

384. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

385. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).

386. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

387. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

388. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4277.

389. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

390. Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).

391. Bishara, A. *et al.* Read clouds uncover variation in complex regions of the human genome. *Genome Res.* **25**, 1570–1580 (2015).

392. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).

393. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

394. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–63 (2014).

395. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).

396. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4**, Unit 4.10 (2009).

397. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).

398. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit - Interactive Quality Assessment of Genome Assemblies. *G3* **10**, 1361–1374 (2020).

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

400. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).

401. Secomandi, S. *et al.* Pangenomics provides insights into the role of synanthropy in barn swallow evolution. *bioRxiv* 2022.03.28.486082 (2022) doi:10.1101/2022.03.28.486082.

402. Secomandi, S. *et al.* The genome sequence of the European nightjar, Caprimulgus europaeus (Linnaeus, 1758). *Wellcome Open Res.* **6**, 332 (2021).

403. Richards, S. It's more than stamp collecting: how genome sequencing can unify biological research. *Trends Genet.* **31**, 411–421 (2015).

404. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).

405. Coster, W. D., De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nature Reviews Genetics* vol. 22 572–587 (2021).

The original artwork of the thesis cover was inspired by a picture from the Instagram profile naturfotografie.hoffmann (on the right).